Aalto University
School of Science
Master's Programme in Life Science Technologies

Janne Myllärinen

Data-driven approach to predict neonatal medical diagnoses

Master's Thesis
Espoo, 27th May 2019

Supervisor: Prof. Simo Särkkä
Advisors: Dr. Jaakko Hollmén
Dr. Ali Bahrami Rad

Aalto University
School of Science

Aalto University
School of Science
Master's Programme in Life Science Technologies

ABSTRACT OF
MASTER'S THESIS

Author:	Janne Myllärinen		
Title:	Data-driven approach to predict neonatal medical diagnoses		
Date:	27th May 2019	**Pages:**	viii + 86
Major:	Complex Systems	**Code:**	SCI3060
Supervisor:	Prof. Simo Särkkä		
Advisors:	Dr. Jaakko Hollmén Dr. Ali Bahrami Rad		

Preterm infants with a very low birth weight are at a great risk of dying or of developing certain life-threatening complications due to their underdevelopment. These critically ill infants are treated at neonatal intensive care units, in which their physiological condition is monitored continuously.

In this thesis, machine learning is applied on the monitored parameter recordings and other patient-specific information from Children's Hospital, Helsinki University Hospital. The purpose is to use binary classifiers to predict neonatal mortality and occurrence of three morbidities: bronchopulmonary dysplasia, necrotising enterocolitis, and retinopathy of prematurity. Majority of the current studies have focused on comparing only a few classifiers. Therefore, a wider comparison of classifier algorithms is performed in this work. In addition to a common measure, the prediction performance is evaluated with two less used measures: F_1 score and area under the precision-recall curve. Additionally, the impact of data preprocessing and feature selection on the prediction result is studied.

The results show large differences in the performance of classifiers. Random forests, k-nearest neighbours, and logistic regression result in the highest F_1 scores. The highest values of area under the precision-recall curve are achieved by random forests along with Gaussian processes. If area under the ROC curve is measured, random forests, Gaussian processes, and support vector machines perform the best.

The monitored physiological parameters are time series and their sampling technique can be altered. This shows only a negligible impact on the results. However, lengthening the monitoring time of physiological parameters to 36–48 hours has a little but positive effect on the results. On the other hand, feature selection has a significant role: birth weight and gestational age are crucial for a high performance. Further, combining them with other features improves the performance. For all that, the optimal data preprocessing procedure is classifier- and complication-specific.

Keywords:	machine learning, binary classification, neonatal complications, prediction
Language:	English

Aalto-yliopisto
Perustieteiden korkeakoulu
Life science -teknologian koulutusohjelma

DIPLOMITYÖN TIIVISTELMÄ

Tekijä:	Janne Myllärinen		
Työn nimi:	Dataan perustuva tapa ennustaa vastasyntyneiden lääketieteellisiä diagnooseja		
Päiväys:	27.5.2018	**Sivumäärä:**	viii + 86
Pääaine:	Monimutkaiset järjestelmät	**Koodi:**	SCI3060
Valvoja:	Prof. Simo Särkkä		
Ohjaajat:	TkT Jaakko Hollmén FT Ali Bahrami Rad		

Syntymäpainoltaan hyvin pienet keskoset ovat suuressa riskissä kuolla tai saada hengenvaarallisia komplikaatioita alikehittyneisyyden takia. Näitä vakavasti sairaita vauvoja hoidetaan vastasyntyneiden teho-osastoilla, joissa heidän fysiologista kuntoaan valvotaan jatkuvasti.

Tämä tutkielma soveltaa koneoppimista valvottujen parametrien tallenteisiin ja muihin potilaskohtaisiin tietoihin, jotka on saatu HUS:n Lastenklinikalta. Tarkoituksena on käyttää binääristä luokittelua ennustamaan vastasyntyneiden kuolleisuutta ja kolmen sairauden puhkeamista. Nämä sairaudet ovat bronkopulmonaalinen dysplasia, nekrotisoiva enterokoliitti sekä keskosten retionopatia. Suurin osa nykyisestä tutkimuksesta on keskittynyt vertailemaan vain muutamia luokittelijoita. Tässä työssä vertaillaan siksi suurempaa määrää eri luokittelualgoritmeja. Yhden yleisesti käytetyn mitan lisäksi ennusteita arvioidaan myös kahdella vähemmän käytetyllä arviointimitalla: F_1-arvolla ja tarkkuus–herkkyys-käyrän alapuolisella alueella. Myös datan esikäsittelyn ja piirteiden valinnan vaikutusta ennustustulokseen tutkitaan.

Tulokset osoittavat suuria eroja eri luokittelijoiden välillä. Satunnaismetsillä, k-lähimmän naapurin luokittimella sekä logistisella regressiolla saadaan korkeimmat F_1-arvot. Suurimmat tarkkuus–herkkyys-käyrän alapuoliset alueet saavutetetaan satunnaismetsillä sekä Gaussisten prosessien luokittimilla. Jos taas ROC-käyrän alapuolinen alue mitataan, satunnaismetsät, Gaussisten prosessien luokitin ja tukivektorikoneet toimivat parhaiten.

Seuratut fysiologiset parametrit ovat aikasarjoja, joten niiden näytteenottotapaa voidaan muuttaa. Tällä on vain pieni vaikutus tuloksiin. Fysiologisten parametrien seuranta-ajan pidentämisellä 36–48 tuntiin on kuitenkin pieni, mutta myönteinen vaikutus tuloksiin. Piirteiden valinnalla on puolestaan merkittävästi väliä: syntymäpaino ja gestaatioikä ovat ratkaisevia hyvien tulosten saamiseksi. Niiden yhdistäminen muiden piirteiden kanssa parantaa tuloksia. Ihanteellinen datan esikäsittely on kaikesta huolimatta luokittelija- ja komplikaatiokohtaista.

Asiasanat:	koneoppiminen, binäärinen luokittelu, vastasyntyneiden komplikaatiot, ennustaminen
Kieli:	Englanti

Acknowledgements

I would like to thank Professor Simo Särkkä for offering me this fascinating Master's thesis position at the edge of data science and medical engineering in the research group of Sensor Informatics and Medical Engineering and for supervising my thesis. In addition, I want to thank my thesis advisors Dr. Jaakko Hollmén and Dr. Ali Bahrami Rad for valuable feedback and support during the thesis process.

I acknowledge Professor Sture Andersson and Dr. Markus Leskinen for introducing me to the world of neonatology, which I was not acquainted with before starting the thesis last autumn, and for valuable advice in medicine related matters. Furthermore, I would like to thank Dr. Olli-Pekka Rinta-Koski for all the practicalities, discussions about the data science in neonatology, and the earlier work that founded an excellent basis for this Master's thesis. I am also grateful to all the competent and splendid colleagues at my own and at the neighbouring research groups for the daily discussions.

And finally, I would like to thank my family and friends at home and abroad for supporting me throughout the years.

Espoo, 27$^{\text{th}}$ May 2019

Janne Myllärinen

Contents

Abbreviations and Acronyms

APACHE	Acute Physiology And Chronic Health Evaluation
AUPR	Area under the precision-recall curve
AUROC	Area under the receiver operating characteristics curve
BPD	Bronchopulmonary dysplasia
BW	Birth weight
const	constant
CRIB	Clinical Risk Index for Babies
DT	Decision tree
EHR	Electronic health records
ELBW	Extremely low birth weight
FN	False negative
FP	False positive
FPR	False positive rate
GA	Gestational age
GP	Gaussian process
HR	Heart rate
HRC	Heart rate characteristics
ICD	International statistical classification of diseases and related health problems
ICU	Intensive care unit
IrregAll	Irregular sampling, all hours included
IrregExcl6h	Irregular sampling, first six hours of life excluded
k-NN	k-nearest neighbours
LDA	Linear discriminant analysis
LOCF	Last-observation-carry-forward
LR	Logistic regression
m32	Matérn kernel with $\nu = 3/2$
m52	Matérn kernel with $\nu = 5/2$
MEWS	Modified Early Warning Score

MIMIC	Multiparameter Intelligent Monitoring in Intensive Care
NB	Naïve Bayes
NEC	Necrotising enterocolitis
NICU	Neonatal intensive care unit
NTISS	National Therapeutic Intervention Scoring System
PAA	Piecewise aggregate approximation
PPV	Positive predictive value
PR	Precision-recall
PRISM	Paediatric Risk of Mortality
QDA	Quadratic discriminant analysis
RBF	Radial basis function
RegAll	Regular sampling, all hours included
RegExcl6h	Regular sampling, first six hours of life excluded
RF	Random forest
ROC	Receiver operating characteristics
ROP	Retinopathy of prematurity
SAPS	Simplified Acute Physiology Score
SE	Standard error
SC	Scores SNAP-II and SNAPPE-II
SNAP	Score for Neonatal Acute Physiology
SNAP-II	Score for Neonatal Acute Physiology II
SNAP-PE	Score for Neonatal Acute Physiology – Perinatal Extension
SNAPPE-II	Score for Neonatal Acute Physiology – Perinatal Extension II
SOFA	Sepsis-related Organ Failure Assessment
SpO_2	Peripheral oxygen saturation
SQL	Structured query language
SVM	Support vector machine
TN	True negative
TNR	True negative rate
TP	True positive
TPR	True positive rate
TS	Time series
VLBW	Very low birth weight
VLGA	Very low gestational age

1. Introduction

Digitalisation of healthcare generates vast amounts of patient-specific medical data. At intensive care units (ICUs), they contain measurement values from patient monitoring, laboratory test results, and clinical notes written by doctors and nurses. These data enable opportunities for machine learning to discover knowledge (Meyfroidt et al., 2009). Various machine learning approaches with various purposes have been proposed to analyse all types of data originated from human beings. They include, but are not limited to, biometric authentication from electroencephalogram signals (Haukipuro et al., 2019), prediction of morbidities associated with preterm birth from physiological parameter measurements (Saria et al., 2010), sequencing genomic data (Libbrecht and Noble, 2015), detection of arrhythmia from electrocardiogram recordings (Suotsalo and Särkkä, 2017), and segmentation of the anatomical regions of the brain from magnetic resonance images (de Brébisson and Montana, 2015).

Physiology of patients is monitored continuously during their stay at ICU which applies also to the smallest patients of all, the preterm infants, which are taken care of at neonatal ICUs (NICUs). These patients are prone to life-threatening complications of preterm birth that are a consequence of their bodies and vital functions not being as developed as those of term infants (McGregor, 2013). Sadly, preterm birth is a major reason for the worldwide mortality of children under the age of five years (WHO and MCEE, 2018). Fortunately, machine learning may provide a solution, or at least a help, when applied on the physiological parameter measurements and other relevant data of preterm infants. Machine learning algorithms may be utilised at NICUs for predicting certain medical complications related to, for instance, respiratory system or sight (McGregor, 2013). Evidence for the applicability of machine learning on the neonatal health care exists. Among others, Ferreira et al. (2012) diagnosed neonatal jaundice from a large number of health-related parameters, Temko et al. (2011) predicted neonatal seizures from electroencephalography data, and Rinta-Koski et al. (2017b, 2018) used several

physiological parameters and other information to predict a few prevalent neonatal morbidities as well as neonatal mortality.

Even though medical doctors are experts in their field, there is a need for data-driven analyses if multiple physiological parameters affect concurrently the well-being and survival of infants. Humans are capable of analysing and recognising patterns from data with three dimensions at most, but we are not able to interpret accurately the data of higher dimensionality (Holzinger, 2016). Accordingly, a computer – together with machine learning algorithms and different types of medical data – is required to perform those analyses. Nonetheless, the intention is not to replace the doctors with algorithms but to provide them with real-time decision support tools. The tools can monitor the patients and suspect potential complications in advance so that doctors can evaluate these patients more carefully (Mani et al., 2014).

During 1999–2013, the NICU at Children's Hospital, Helsinki University Hospital has been collecting and storing masses of data for more than 2,000 preterm infant patients with a very low birth weight (VLBW). This number corresponds to around one-third of all Finnish VLBW infants born during those years. This database is exceptionally wide in terms of temporal scale and coverage, also globally. A few studies, including Immeli et al. (2017) and Rinta-Koski et al. (2017b), have already utilised this database.

A decent amount of research has been conducted on predicting medical complications with machine learning algorithms. However, most of those studies have repeatedly applied the same algorithms to make predictions, and the literature is lacking their wider comparison. Therefore, the first research objective of this study is to determine which algorithms are the most suitable for predicting neonatal complications and if there are differences in the predictability of different complications. This is executed by applying 12 machine learning algorithms on neonatal mortality and three morbidities, and by comparing their predictive capabilities.

Patient cohorts are often imbalanced, meaning the ratio of sick patients to all subjects is low. Due to the rareness of sick patients, identifying them is challenging from the machine learning point of view. If machine learning algorithms are applied on imbalanced data and evaluated inappropriately, they tend to show misleading results. This is the case in many of the previous studies. They evaluate the results using accuracy and area under the receiver operating characteristics curve (see Section 2.4) and receive questionably high results (Saito and Rehmsmeier, 2015; Rokach, 2010; Rollins et al., 2015; Libbrecht and Noble, 2015). Using incorrect measures can have fatal consequences if the sick patients are not identified and given medical treat-

ment on time, but the measure still shows a high performance. Therefore, the second goal of this work is to present less-used measures that function more truthfully with imbalanced data. These measures and a more commonly used measure are applied to evaluate the performance of machine learning algorithms. Further, the results of this work are compared to previous studies. Since making reliable comparisons between distinct datasets is challenging (Salcedo-Bernal et al., 2016), the results are primarily compared to studies that have been performed on the exactly same neonatal data from the NICU at Helsinki University Hospital.

As the high-quality database has a wide coverage of different types of patient-specific data, the third and more technical research objective of this work is to specify the optimal data preprocessing and feature selection technique for neonatal mortality and morbidity predictions. To be precise, the optimal time series sampling of the temporal physiological parameters and the optimal length of the monitoring time of the same parameters are examined in the preprocessing phase. Moreover, including the most relevant features in the model can improve its prediction performance (Guyon and Elisseeff, 2003). Therefore, the optimal combination of health-related parameters is studied in the feature selection phase.

By finding the best machine learning algorithms, by assessing the results with appropriate evaluation criteria, and by determining the optimal preprocessing and feature selection procedure, the analysis tool could be implemented in real hospital environment some day. This decision support tool would assist medical doctors to plan the treat of the critically ill preterm infants before the complications have occurred or their symptoms become too severe. Foremost, this would improve the care of the neonates, prevent them from developing critical and life-long complications, and save human lives.

The work is structured as follows. Chapter 2 presents the theoretical background, concentrating on data science, and a literature review considering previous studies. Chapter 3 describes the preterm infant data and the methodology how the data have been analysed, followed by the results in Chapter 4. The results are interpreted and the research questions are answered in Chapter 5. Finally, Chapter 6 concludes the work.

2. Background

2.1 Neonatology

The term *neonatology*, a subspecialty of *paediatrics*, has been introduced for the first time in 1960, and it focuses on the medical care and treatment of human newborns, neonates (Avery et al., 2005). This section provides a brief introduction to neonates, their medical complications, patient monitoring, and traditional scores to evaluate patients' physical condition.

2.1.1 Neonatal infants

Neonates, which require critical care at neonatal intensive care units, are most often preterm infants, who are prone to numerous complications and illnesses due to their underdeveloped organs and young age (McGregor, 2013; Avery et al., 2005). Approximately 15 million preterm infants are born worldwide annually, which corresponds to more than 10 % of all neonates, but this rate, however, varies country-specifically between 5 % and 18 % (WHO, 2018).

Gestational age (GA) and birth weight (BW) are important and widely used attributes to describe neonates. GA means the time period from the first day of the last normal menstrual period of the mother to the day of delivery, and GA is usually reported in weeks (American Academy of Pediatrics, 2004). If GA of a newborn is less than 32 weeks, the infant is said to have a very low gestational age (VLGA) (Fattore et al., 2015). In addition, infants born before the gestational age of 37 weeks are called preterm, between the 37^{th} and the 41^{st} week are term, and after the 41^{st} week are post term (Gomella et al., 2013). Very low birth weight (VLBW) infants weigh less than 1500 g, and extremely low birth weight (ELBW) infants less than 1000 g (Avery et al., 2005; Gomella et al., 2013).

2.1.2 Typical neonatal complications

ELBW infants tend to have all kinds of health issues that can be respiratory (e.g., respiratory distress syndrome), cardiovascular (e.g., patent ductus arteriosus), central nervous system (e.g., intraventricular haemorrhage), renal (e.g., electrolyte imbalance), ophthalmologic (e.g., retinopathy of prematurity), gastrointestinal–nutritional (e.g., necrotising enterocolitis or jaundice), or immunologic (e.g., proneness to infections) problems (Avery et al., 2005). Critical care of VLBW and VLGA infants is costly, and according to Fattore et al. (2015), the cost of saving one preterm infant from very likely death is €20,000–€40,000. In this study, neonatal mortality as well as bronchopulmonary dysplasia, necrotising enterocolitis, and retinopathy of prematurity are of a special interest.

Neonatal mortality has been on a decrease during the ongoing millennium as Figure 2.1 presents (United Nations, 2019). Still, it corresponds to 2.5 million annual deaths globally (UNICEF et al., 2018). Complications of preterm birth caused almost 0.9 million of all neonatal deaths, which also accounts for approximately 6 % of all 15 million annually born preterm infants (WHO and MCEE, 2018; WHO, 2018). What is more, the mortality rate among VLBW and VLGA infants is even higher. In Finland, it is 11.4 % one month after the birth and 11.7 % after one year (Fattore et al., 2015).

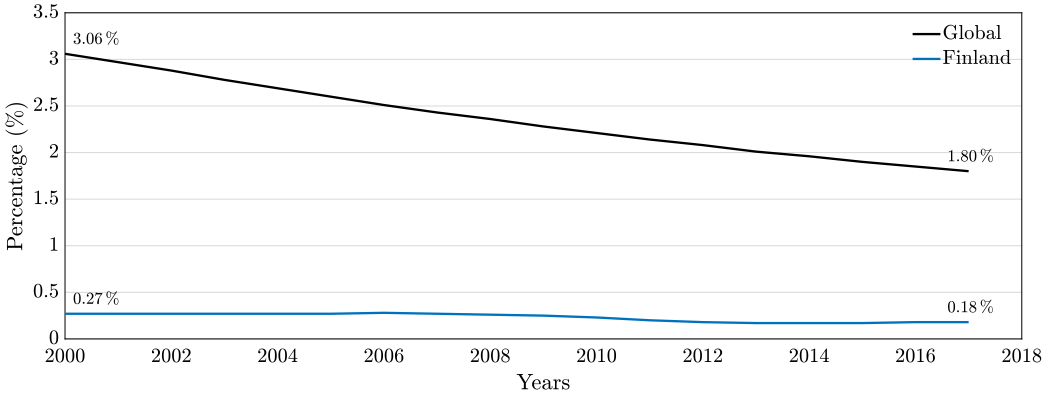


Figure 2.1: Neonatal mortality rate globally and in Finland during 2000–2017. Data from United Nations (2019).

Bronchopulmonary dysplasia (BPD) is a chronic lung disease, developed in preterm infants due to factors compromising normal development in the immature lung, such as treatment of additional oxygen and the use of mechanical ventilation (Avery et al., 2005). A low birth weight and gestational age are associated with the risk of developing BPD (Gomella et al., 2013;

Wajs et al., 2006, 2007). Approximately 30 % of ELBW infants are diagnosed with BPD (Gomella et al., 2013; Walsh et al., 2006).

Necrotising enterocolitis (NEC) is a disease of gastrointestinal tract of preterm neonates, where inflammation and bacterial invasion of the bowel wall leads to necrosis. Around 6 %–10 % of VLBW infants have NEC, and the more preterm infants are at a higher risk of NEC (Gomella et al., 2013).

Retinopathy of prematurity (ROP) is a maldevelopment of the retinal vasculature, caused by interrupted retinal vessel formation, the symptoms of which vary in severity and can lead to blindness at worst (Gomella et al., 2013). Supplementary oxygen given to infants is often believed to contribute to the development of ROP (Cirelli et al., 2013; Gomella et al., 2013). In addition, a low birth weight correlates with the rate of developing ROP (Darlow et al., 2005). To prevent ROP, controlling and optimising the oxygen saturation of the patient is essential as well as maintaining the physiological state of the patient stable to avoid infections, and thus, abnormal growth and development of the patient (Hellström et al., 2013).

2.1.3 Evaluating the neonatal condition

Throughout the years, several scoring systems have been introduced to numerically evaluate the condition of newborn infants. Demographic, physiological, and clinical data are used to calculate the scores, which give mortality and different morbidities a quantification and are used to identify the high-risk patients (Dorling et al., 2005). Two types of scores exist: medical and statistical. Medical experts have defined the parameters and their weights used in medical scores, whereas the statistically relevant parameters have been selected for statistical scores (Dorling et al., 2005). The medical scores are easier to be understood by the personnel using them, but their disadvantage is the worse performance in comparison to the statistical scores.

Multiple medical scores are discussed in the literature. National Therapeutic Intervention Scoring System, NTISS, is calculated from 62 values and used to predict mortality and assess severity of illnesses (Gray et al., 1992). The Apgar score evaluates the neonatal condition from five signs (Apgar, 1953). The illness severity index and predictor of mortality Score for Neonatal Acute Physiology, SNAP, is calculated from 34 values for VLBW infants (Richardson et al., 1993a). Its extension, Score for Neonatal Acute Physiology – Perinatal Extension, SNAP-PE, is calculated from SNAP and three additional values using logistic regression (Richardson et al., 1993b).

Statistical techniques have been applied to select the parameters for the simplified versions of SNAP and SNAP-PE, namely SNAP-II and SNAPPE-II (Richardson et al., 2001). SNAP-II is calculated from six values and SNAPPE-II from SNAP-II and three additional values, which are similar to those of SNAP-PE (Richardson et al., 2001).

Logistic regression has been used to define the parameters for several statistical scores. Clinical Risk Index for Babies, CRIB, predicts mortality for VLBW infants or infants with GA of less than 31 weeks from six values (International Neonatal Network, 1993). Its simplified version, CRIB II, is calculated from five redefined values for neonates with GA of 32 weeks (Parry et al., 2003). Berlin score (Maier et al., 1997) uses five values to assess the mortality risk of VLBW patients.

Additionally, many other scores exist, and they evaluate the condition of child and adult patients. They include, but are not limited to, Acute Physiology And Chronic Health Evaluation, APACHE, (Knaus et al., 1981) along with the revised versions APACHE II (Knaus et al., 1985), APACHE III (Knaus et al., 1991), and APACHE IV (Zimmerman et al., 2006), Glasgow Coma Score (Teasdale and Jennett, 1974), Modified Early Warning Score, MEWS, (Subbe et al., 2001), Pediatric Risk of Mortality, PRISM, (Pollack et al., 1988) with its revised version PRISM III (Pollack et al., 1996), Simplified Acute Physiology Score, SAPS, (Le Gall et al., 1984) and its revised version SAPS II (Le Gall et al., 1993) as well as Sepsis-related Organ Failure Assessment, SOFA, (Vincent et al., 1996) and quickSOFA (Singer et al., 2016).

Even though certain scores are widely adopted and used for research purposes, a single score cannot explain the true condition of an infant as they always emphasise some aspects over others (Dorling et al., 2005). The use of scores has also been criticised as they are static values, calculated at single time points only, and are not updated over time (Ghassemi et al., 2015). Therefore, continuous patient monitoring is essential in gaining correct information about the condition of the patients.

2.1.4 Monitoring the neonatal physiological variables

The human physiology is monitored with various sensors to have an updated view on the patient's condition so that potential onset of medical complications can be prevented by intervening them in advance (Murković et al., 2003). At NICUs, the infants are kept in incubators, where the temperature and humidity conditions are appropriate. What is more, multiple functionalities are integrated into incubators which can be medical care devices, such as

ventilators, or patient monitoring devices, such as pulse oximetry. The monitored parameters usually include, but are not limited to, electrocardiography, electroencephalography, heart rate (HR), blood pressure, temperature, respiratory rate, and peripheral blood oxygen saturation (SpO_2) (Rinta-Koski, 2018; Murković et al., 2003).

The measurements quantify the state of preterm infant patients, which is a requirement for machine learning applications. Thus, the measurements form the integral basis for this study since the continuous parameter monitoring enables to evaluate and model the patient's condition with machine learning algorithms instead of static scores.

2.2 Time series analysis

This section introduces time series and describes how information can be extracted from them. Furthermore, techniques to identify the relevant features from all possible features are discussed in Section 2.2.3.

2.2.1 Time series

A time series consists of multiple consecutive observations of a parameter, measured over a certain time period (Batal et al., 2009). Each observation has a value and a corresponding time stamp. If multiple parameters are measured simultaneously, the time series is called multivariate.

Similar temporal patterns are searched from physiological time series as they may correspond to certain clinical diagnoses (Lehman et al., 2008). Consequently, the appearance of these patterns can reveal upcoming complications before the condition of the patient deteriorates. Using time series and more complex temporal information may improve the prediction performance. Temporal patterns may include information that is not visible from a single value; relationships between certain parameters and medication intake can contain more information than only the newest monitored parameter values (Batal et al., 2009).

2.2.2 Feature extraction

Feature extraction means finding the essential information from, potentially, massive amounts of data usually by reducing the dimensionality of the data and by compressing the data into features (Duda et al., 2001). Based on the features, dissimilar data can be distinguished from each other. Time series features include, for example, regression slopes in certain intervals, maximum transient increase and decrease of the values, and similarity measures within and between signals, of which autocorrelation coefficients measure the within-signal similarity and cross correlation coefficients the between-signal similarity (Lehman et al., 2008). Autoregressive–moving-average parameters, introduced by Wold (1938), are also a technique to extract information from time series.

Temporal abstraction patterns can be extracted from time series data using four methods as follows (Batal et al., 2012).

1. *Temporal abstractions* transform raw, multivariate time series data into a symbolic form where information is encoded to a higher abstraction level (Moskovitch and Shahar, 2015). They are divided into two methods:
   (a) *value* or *state abstractions* categorise values to groups, such as low, normal, and high, and
   (b) *trend abstractions* categorise time intervals of predefined length to groups, such as increasing, steady, and decreasing (Batal et al., 2009; Sacchi et al., 2007).
2. *Multivariate state sequences* observe the value abstraction sequences over time for multiple time series.
3. *Temporal relations* are based on Allen's temporal logic (Allen, 1984), and they observe the timing of the occurrence of certain events, for example, consecutive occurrences, or partly or totally overlapping occurrences.
4. *Temporal patterns* observe the sequence of temporal relations.

Shapelets are another technique to extract information from time series. They are defined as exceptionally representative subsequences of the class, in which the whole time series belongs to (Ye and Keogh, 2009). In other words, shapelets find the relevant parts of time series that include enough information to classify the whole time series. One more algorithm to identify temporal patterns is segmented time series feature mine (Batal et al., 2009), which is based on the Apriori algorithm by Agrawal and Srikant (1994).

2.2.3 Feature selection

The number of extractable features is enormous. In feature selection, the number of extracted features is reduced so that only the most relevant ones are used in classification (Murphy, 2012). This improves the performance of the prediction, makes the computation more efficient, and explains what is essential in the underlying data (Guyon and Elisseeff, 2003; Salcedo-Bernal et al., 2016). However, Temko et al. (2011) prefer including all available features for support vector machine classification (see Section 2.3.9) since the presence of redundant features does not distract the classifier, unlike the lack of important features. As an acknowledgement, a variable, which does not improve the classification result alone, can improve it together with other variables (Guyon and Elisseeff, 2003).

Three common feature selection techniques are filter, wrapper, and embedded methods, for which the reader is advised to refer to Guyon and Elisseeff (2003). *Filter methods*, such as the correlation criterion of the square of Pearson correlation coefficient, are suitable for binary classification. For instance, features with the lowest correlation with the outcome variable can be omitted from the model, which, however, may simultaneously decrease the classification result (Salcedo-Bernal et al., 2016). *Wrapper methods* apply the machine learning algorithm of interest to identify the optimal features. They either select, as in *forward selection*, or omit, as in *backward elimination*, the features one by one, ending up to a locally optimal performance. *Embedded methods* are a combination of filter and wrapper methods that can improve the results in comparison to filter methods, but the improvement is not guaranteed to be significant.

2.3 Machine learning classification methods

This section presents the principles of machine learning with a focus on describing how classifiers determine the class for data points.

2.3.1 Machine learning and classification in general

A high-level division of machine learning is supervised and unsupervised learning (Hastie et al., 2001; Goodfellow et al., 2016; Murphy, 2012). The goal in both of them is to build a *model* that discovers knowledge from data, which are split into *training data* and *test data*. The training data

consist of input-output pairs $\mathcal{D} = \{(\boldsymbol{x}^{(i)}, y^{(i)})\}_{i=1}^{N}$, where N is the number of *observations* (also called as *data points*, *data instances*, or *cases*), each of which is required to have d known *features* $\boldsymbol{x}^{(i)} = (x_1^{(i)}, \ldots, x_d^{(i)})$ (also called as *attributes*, *predictive attributes*, or *explanatory variables*) and one possibly known *outcome variable* $y^{(i)}$ (also called as *class*, *label*, *target*, or *response variable*) (Bellazzi and Zupan, 2008; Hastie et al., 2001; Goodfellow et al., 2016; Murphy, 2012; Bishop, 2006).

In *supervised learning*, the known features $\boldsymbol{X} = (\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(N)})^T$ and the known outcome variables $\boldsymbol{y} = (y^{(1)}, \ldots, y^{(N)})^T$ of the training data are used to build a model. The purpose of the model is to predict the unknown outcome variable y of unseen data instance of test data from their known features $\boldsymbol{x}$ by estimating the probability $p(y \mid \boldsymbol{x})$ (Lucas, 2004; Hastie et al., 2001; Goodfellow et al., 2016). If the outcome variable can only have discrete values or is qualitative, the machine learning problem is called *classification*, whereas a continuous outcome variable implies *regression* (Meyfroidt et al., 2009; Hastie et al., 2001).

In *unsupervised learning*, on the other hand, the outcome variables $\boldsymbol{y}$ are unknown, and the aim is to observe the features in the unlabeled data $\mathcal{D} = \{(\boldsymbol{x}^{(i)})\}_{i=1}^{N}$ to learn the probability distribution $p(\boldsymbol{x})$ (Murphy, 2012). The model is build by finding certain patterns in the attributes, based on which certain data points are grouped or *clustered* together (Meyfroidt et al., 2009). In addition to clustering, unsupervised learning covers, for example, *association rules* and *self-organising maps* (Hastie et al., 2001).

In the ICU context, an interesting question is to predict the survival of patients, which can be implemented as a supervised binary classification problem (Meyfroidt et al., 2009). In classification, the purpose is to build a model based on the training data, and then generalise the model on unseen data instances. The features of an unseen data point $\boldsymbol{x}$ are used to assign the data point with a label $y \in \{C_1, \ldots, C_K\}$ that represents one of K discrete classes C_k, where $k = 1, \ldots, K$ (Bishop, 2006). The classes are separated by *decision boundaries*, also known as *decision surfaces*, from each other in the feature space.

In this work, the data instances are NICU patients and the input data consist of their physiological parameter measurements and other patient-specific information. Furthermore, the outcome variable $y^{(i)} \in \{0, 1\}$, where $y^{(i)} = 0$ denotes the class C_1, the patient i dies or is given a certain diagnosis, and $y^{(i)} = 1$ denotes the class C_2, the patient i does not die or is not given the diagnosis.

The generalisation capability is measured by *generalisation error*, *test error*

or *classification error*, which means the probability to misclassify an unseen data instance from the test data (Goodfellow et al., 2016; Rokach, 2010). Additionally, the machine learning models are evaluated with *training errors* which are errors due to misclassification, calculated from the training set. Minimising the training error means optimising the parameters of the model for the training set so accurately that the generalisation capability of the model is reduced (Goodfellow et al., 2016). Thus, the test error increases, which is referred to as *overfitting*. It is one of the major challenges in machine learning.

In the field of medicine, the most widely used machine learning classifiers include decision trees, random forests, artificial neural networks, Bayesian networks, support vector machines, and Gaussian processes, and there is no evidence that a certain classifier would be more suitable for a certain task than any other (Meyfroidt et al., 2009). Therefore, a variety of classifiers are applied and compared to determine the most suitable classifiers for neonatal complication predictions to respond to the first research objective of this work.

2.3.2 Gaussian processes

Gaussian processes (GPs) are generalisations of the Gaussian probability distribution, and they belong to probabilistic classification methods that produce probabilities of belonging to a class instead of bare class labels (Bishop, 2006; Rasmussen and Williams, 2006). The goal of Gaussian processes is to learn the distribution over function for the given data $p(f \mid \boldsymbol{X}, \boldsymbol{y})$, and then determine the *posterior* or *predictive probability* $p(y^* \mid \boldsymbol{X}, \boldsymbol{y}, \boldsymbol{x}^*)$ to predict the label y^* for a test data point $\boldsymbol{x}^*$ (Rasmussen and Williams, 2006; Murphy, 2012). An example of GP classification result is presented in Figure 2.2. Next, binary GP classification is described in more detail, and the test data point is denoted with an asterisk (*) for clarity.

First, a Gaussian process prior is adapted over a latent function $\boldsymbol{f}^* = (f(\boldsymbol{x}^{(1)}), \ldots, f(\boldsymbol{x}^{(N)}), f(\boldsymbol{x}^*))$, which is defined as in Equation (2.1),

$$p(\boldsymbol{f}^*) = \mathcal{N}(\boldsymbol{f}^* \mid \boldsymbol{0}, \boldsymbol{\Sigma}^*), \tag{2.1}$$

where the covariance matrix $\boldsymbol{\Sigma}^*$ consists of elements $\Sigma(\boldsymbol{x}, \boldsymbol{x}') = k(\boldsymbol{x}, \boldsymbol{x}')$, in which $k(\boldsymbol{x}, \boldsymbol{x}')$ is any positive semidefinite kernel function (Bishop, 2006; Murphy, 2012). For a test data point, the distribution of this latent variable f^* is defined by Rasmussen and Williams (2006) as in Equation (2.2),

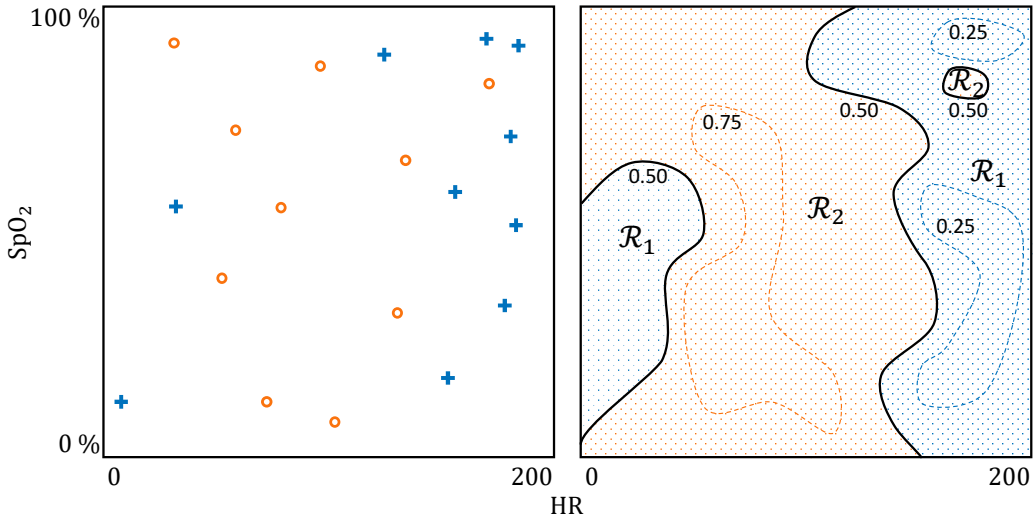


Figure 2.2: A possible result of a GP classification based on two features: heart rate (HR) and peripheral oxygen saturation (SpO_2). The left part shows the locations of the data points of the blue and orange classes, and the right part shows the contour plots for the predictive probabilities, where the black line represents the decision boundary between decision regions $\mathcal{R}_1$ (blue class) and $\mathcal{R}_2$ (orange class). Figure following Rasmussen and Williams (2006).

$$p(f^* \,|\, \boldsymbol{X}, \boldsymbol{y}, \boldsymbol{x}^*) = p(f^* \,|\, \boldsymbol{X}, \boldsymbol{y}, \boldsymbol{f})\; p(\boldsymbol{f} \,|\, \boldsymbol{X}, \boldsymbol{y})\; d\boldsymbol{f}. \tag{2.2}$$

Second, a logistic sigmoid function $\sigma(f^*) = (1 + \exp(f^*))^{-1}$ is applied on the latent to transform the result from the whole span of the x-axis into the interval of $[0, 1]$ to receive an appropriate binary classification result (Bishop, 2006; Rasmussen and Williams, 2006).

Third, it is sufficient to calculate the posterior distribution only for one class $p(y^* = 1 \,|\, \boldsymbol{X}, \boldsymbol{y}, \boldsymbol{x}^*)$ since the posterior distribution for the other class is simply its complement $p(y^* = 0 \,|\, \boldsymbol{X}, \boldsymbol{y}, \boldsymbol{x}^*) = 1 - p(y^* = 1 \,|\, \boldsymbol{X}, \boldsymbol{y}, \boldsymbol{x}^*)$. Following Bishop (2006) and Rasmussen and Williams (2006), the probabilistic prediction is calculated as a combination of the previous steps as in Equation (2.3),

$$p(y^* = 1 \,|\, \boldsymbol{X}, \boldsymbol{y}, f^*) = \int \sigma(f^*)\; p(f^* \,|\, \boldsymbol{X}, \boldsymbol{y}, \boldsymbol{x}^*)\; df^*. \tag{2.3}$$

Kernel functions for Gaussian processes

The choice of the *covariance matrix* or *kernel function* $\mathbf{\Sigma}$ is essential in GP classification since assumptions of the similarities between data points are encoded in that (Rasmussen and Williams, 2006). Different kernels include, but are not limited to, constant, linear, squared exponential or radial basis function (RBF), and Matérn kernels in Equations (2.4a), (2.4b), (2.4c), and (2.4d), respectively,

$$k_{\text{const}}(\boldsymbol{x}, \boldsymbol{x}') = \sigma^2, \tag{2.4a}$$

$$k_{\text{linear}}(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{x}^T \mathbf{\Sigma} \boldsymbol{x}', \tag{2.4b}$$

$$k_{\text{RBF}}(\boldsymbol{x}, \boldsymbol{x}') = \exp(-\frac{r^2}{2\ell^2}), \tag{2.4c}$$

and

$$k_{\text{Matérn}}(\boldsymbol{x}, \boldsymbol{x}') = \frac{1}{2^{\nu-1}\Gamma(\nu)} \left(\frac{\sqrt{2\nu}}{\ell} r \right)^{\nu} K_\nu \left(\frac{\sqrt{2\nu}}{\ell} r \right), \tag{2.4d}$$

where $\boldsymbol{x}$ and $\boldsymbol{x}'$ are a pair of inputs, σ^2 is a variance, $r = \boldsymbol{x} - \boldsymbol{x}'$ is a stationary covariance function, ℓ is a characteristic length-scale, ν is a positive parameter, K_ν is a modified Bessel function (see Abramowitz and Stegun (1965)), and Γ is the gamma function (Rasmussen and Williams, 2006; Murphy, 2012). According to Rasmussen and Williams (2006), the most interesting Matérn kernels from the machine learning perspective are the ones with parameters $\nu = 3/2$ and $\nu = 5/2$ as in Equations (2.4e), and (2.4f),

$$k_{\text{Matérn32}}(\boldsymbol{x}, \boldsymbol{x}') = \left(1 + \frac{\sqrt{3}r}{\ell} \right) \exp \left(-\frac{\sqrt{3}r}{\ell} \right), \tag{2.4e}$$

and

$$k_{\text{Matérn52}}(\boldsymbol{x}, \boldsymbol{x}') = \left(1 + \frac{\sqrt{5}r}{\ell} + \frac{5r^2}{3\ell^2} \right) \exp \left(-\frac{\sqrt{5}r}{\ell} \right), \tag{2.4f}$$

respectively.

Valid kernels can be constructed from other valid kernels by following simple rules (Bishop, 2006). For example, a sum or a product of two valid kernels results in a valid kernel (Rasmussen and Williams, 2006). In this work, four

distinct kernels are applied, and they correspond to the kernels of Rinta-Koski et al. (2018). These kernels are sums of linear, constant, and kernel-specifically optionally one of the kernels presented above, and the constructed kernels are as in Equations (2.5a), (2.5b), (2.5c), and (2.5d),

$$k_1(\boldsymbol{x}, \boldsymbol{x}') = k_{\text{linear}}(\boldsymbol{x}, \boldsymbol{x}') + k_{\text{const}}(\boldsymbol{x}, \boldsymbol{x}'), \tag{2.5a}$$
$$k_2(\boldsymbol{x}, \boldsymbol{x}') = k_{\text{linear}}(\boldsymbol{x}, \boldsymbol{x}') + k_{\text{const}}(\boldsymbol{x}, \boldsymbol{x}') + k_{\text{Matérn32}}(\boldsymbol{x}, \boldsymbol{x}'), \tag{2.5b}$$
$$k_3(\boldsymbol{x}, \boldsymbol{x}') = k_{\text{linear}}(\boldsymbol{x}, \boldsymbol{x}') + k_{\text{const}}(\boldsymbol{x}, \boldsymbol{x}') + k_{\text{Matérn52}}(\boldsymbol{x}, \boldsymbol{x}'), \tag{2.5c}$$

and

$$k_4(\boldsymbol{x}, \boldsymbol{x}') = k_{\text{linear}}(\boldsymbol{x}, \boldsymbol{x}') + k_{\text{const}}(\boldsymbol{x}, \boldsymbol{x}') + k_{\text{RBF}}(\boldsymbol{x}, \boldsymbol{x}'). \tag{2.5d}$$

2.3.3 Naïve Bayes

The naïve Bayes classification (NB) is based on the Bayes formula in Equation (2.6),

$$P(y = C_k \,|\, \boldsymbol{x}) = \frac{p(\boldsymbol{x} \,|\, C_k)\, P(C_k)}{p(\boldsymbol{x})}, \tag{2.6}$$

where C_k represents the k^{th} class label, and the posterior probability $P(y = C_k \,|\, \boldsymbol{x})$ for an unknown data instance $\boldsymbol{x}$ is calculated from likelihood $p(\boldsymbol{x} \,|\, C_k)$, prior probability $P(C_k)$, and evidence $p(\boldsymbol{x})$ (Duda et al., 2001; Mitchell, 1997).

The goal of the naïve Bayes classifier is to calculate the maximum posterior probability, and thereby, classify the unseen data point to the most likely class (Duda et al., 2001). Additionally, the denominator in Equation (2.6) is irrelevant under the assumption of conditionally independent features x_j, and it is omitted. The formula simplifies to Equation (2.7),

$$P(y = C_k \,|\, \boldsymbol{x}) = \underset{C_k \in K}{\operatorname{argmax}}\, P(C_k) \prod_{j=1}^{d} p(x_j \,|\, C_k), \tag{2.7}$$

where d is the dimensionality of the feature vector $\boldsymbol{x}$. Additional data instances contribute positively to the performance of the model as they make the posterior probability density function sharper (Duda et al., 2001).

2.3.4 Linear discriminant analysis

Linear discriminant analysis (LDA) divides the d-dimensional space $\mathbb{R}^d$ into classes by hyperplanes whose decision boundaries are linear (Hastie et al., 2001). The decision boundary divides the feature space into two subspaces or *decision regions* $\mathcal{R}_1$ for $y = 0$ and $\mathcal{R}_2$ for $y = 1$ in binary classification (Duda et al., 2001). An example of binary LDA classification is in Figure 2.3(a).

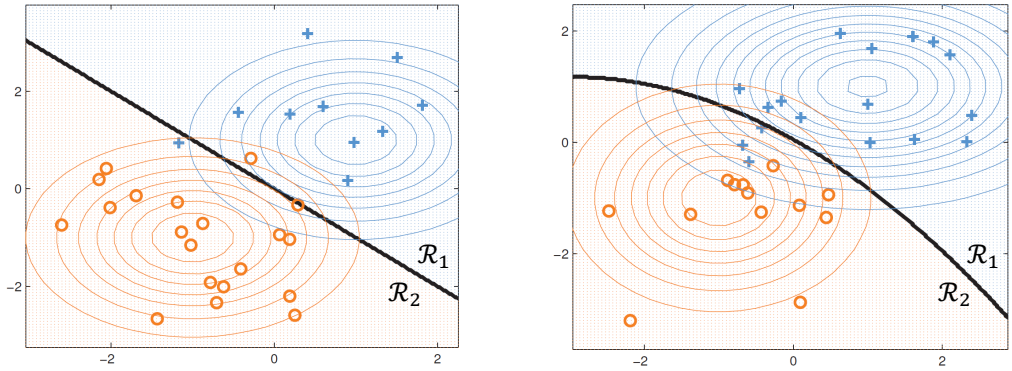
LDA models the class conditional densities as Gaussian distributions as in Equation (2.8),

$$p(\boldsymbol{x} \mid y = C_k, \boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{2.8}$$

where $\boldsymbol{\theta}$ refers to the parameters of the model: the d-dimensional, class-specific mean vector $\boldsymbol{\mu}_k$, and the class-specific covariance matrix $\boldsymbol{\Sigma}_k$ (Murphy, 2012). LDA assumes that all classes have a common covariance matrix (Hastie et al., 2001; Murphy, 2012). Thus, the class-specific covariance matrices simplify to a common covariance matrix as $\boldsymbol{\Sigma}_k = \boldsymbol{\Sigma} \, \forall k$. The posterior probabilities for class labels are formulated as in Equation (2.9),

$$p(y = C_k \mid \boldsymbol{x}, \boldsymbol{\theta}) = \boldsymbol{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k - \frac{1}{2} \boldsymbol{\mu}_k^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k + \log \, \pi_k, \tag{2.9}$$

where π_k denotes the class-specific prior probability $P(C_k)$ (Hastie et al., 2001; Murphy, 2012).



(a) Binary LDA classification with the linear decision boundary

(b) Binary QDA classification with the quadratic decision boundary

Figure 2.3: Two binary discriminant analysis classifiers separate the blue and orange classes. Figure modified from Murphy (2012).

2.3.5 Quadratic discriminant analysis

The linear decision boundaries of LDA (see Section 2.3.4) are not always adequate to separate the classes from each other, and in those cases, quadratic discriminant analysis (QDA) might result in a better classification. QDA has quadratic decision boundaries instead of linear, and the class-specific covariance matrices are not assumed to be equal (Hastie et al., 2001). Thus, each class C_k has its own covariance matrix $\boldsymbol{\Sigma}_k$. Quadratic discriminant functions are formulated as in Equation (2.10),

$$p(y = C_k \,|\, \boldsymbol{x}, \boldsymbol{\theta}) = -\frac{1}{2} \log |\boldsymbol{\Sigma}_k| - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_k) + \log\, \pi_k, \quad (2.10)$$

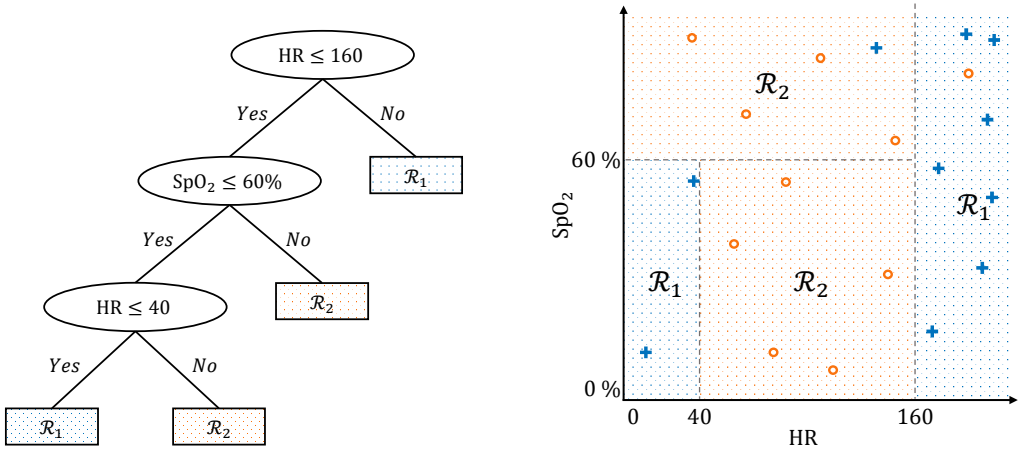
(Hastie et al., 2001). A possible classification of QDA is presented in Figure 2.3(b).

2.3.6 Decision trees

Classification and regression trees have been introduced by Breiman et al. (1984), who recognised their applicability to medical diagnosis predictions. In decision trees (DT), thresholds are set for the feature values, and each threshold splits the data into two non-overlapping subsets $\mathcal{R}_1$ and $\mathcal{R}_2$ at *decision points* (Breiman et al., 1984; Goodfellow et al., 2016; Murphy, 2012). Figure 2.4(a) presents the tree-like structure of DT. More mathematically, the decision points divide the feature space into regions with hyperplanes, resulting in hyper-rectangles that correspond to the leaf nodes (Podgorelec et al., 2002). Rectangle regions $\mathcal{R}_1$ and $\mathcal{R}_2$ are illustrated in Figure 2.4(b).

Splitting the data into smaller subsets is repeated until almost all data instances of the subsets or *leaf nodes* belong to the same class C_k (Goodfellow et al., 2016; Duda et al., 2001). Had all data instances in a leaf node the same outcome variable, the model could be overfitting (Podgorelec et al., 2002). Overfitting is prevented by pruning, in which the number of splits is limited (Murphy, 2012). To test the performance of a decision tree, the features of a new data instance are compared to the thresholds in the tree-like structure, and the label of the leaf node becomes the class of the test data instance.

The advantage of decision trees is the easily understandable rules (Mani et al., 2014; Duda et al., 2001). Decision trees accept both continuous and discrete data as input (Murphy, 2012). They are also relatively robust classifiers to



(a) The logic of a decision tree with decision points (ellipses) and leaf nodes (rectangles)

(b) The same decision tree as regions in feature space

Figure 2.4: A possible result of a decision tree of two features: heart rate (HR) and peripheral oxygen saturation (SpO_2).

labelling errors, and outliers (Meyfroidt et al., 2009; Murphy, 2012). The disadvantages of decision trees include their poor performance on incomplete data, the lack of alternative solutions as they are able to produce only one model for a given problem, and their incapability to emphasise the more important decisions over the less important ones (Podgorelec et al., 2002).

2.3.7 Random forests

Random forests (RF) consist of an ensemble of trees, each of which has been trained with a slightly dissimilar subset of the training data (Murphy, 2012; Meyfroidt et al., 2009). The sampling of the subsets is independent and identically distributed, resulting in slightly dissimilar trees for each sampling (Breiman, 2001). After the trees are grown, their results are averaged or the most common result is voted to be the result of RF model. Accordingly, the model has a lower variance than single decision trees. The number of trees in the forest is not relevant as the generalisation error of the model converges as long as there are sufficiently many trees (Breiman, 2001).

2.3.8 Logistic regression

Despite its name, logistic regression (LR) is a classification method, whose origin lays in linear regression (Bishop, 2006; Goodfellow et al., 2016). Logistic regression models the posterior probabilities of the perfectly separable classes with linear functions (Hastie et al., 2001). The regression coefficients or *weights* $\boldsymbol{w}$ of the functions do not have a closed-form solution but they are optimised with algorithms such as maximum likelihood estimation or gradient descent (Murphy, 2012). Logistic regression is presented in Equation (2.11),

$$p(y = C_k \,|\, \boldsymbol{x}, \boldsymbol{W}) = \frac{\exp(\boldsymbol{w}_k^T \boldsymbol{x})}{\sum_{i=1}^{K} \exp(\boldsymbol{w}_i^T \boldsymbol{x})}, \tag{2.11}$$

where $\boldsymbol{W}$ contains all the class-specific weight vectors $\boldsymbol{w}_k$, and K is the number of classes C_k (Murphy, 2012).

2.3.9 Support vector machines

Support vector machines (SVMs) are generalisations of logistic regression since perfect linear separability of the classes is not required (Hastie et al., 2001). Moreover, SVMs output only the class labels, not the probabilities as LR does (Goodfellow et al., 2016).

SVMs are based on mapping the input data into a high-dimensional feature space where the optimal linear decision boundaries or hyperplanes are set between the classes, so that the margin between the vectors of the classes is maximised (Cortes and Vapnik, 1995). Mathematically, maximising the margin equals to minimising the weight vector $\|\boldsymbol{w}\|^2$, since the margin equals to $\frac{2}{\|\boldsymbol{w}\|}$ (Cortes and Vapnik, 1995; Hastie et al., 2001; Bishop, 2006). Thus, the optimisation problem is as in Equation (2.12),

$$\begin{aligned} &\min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|^2 \\ &\text{subject to } y^{(i)}(\boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}^{(i)}) + b) \geq 1, \qquad i = 1, \ldots, N, \end{aligned} \tag{2.12}$$

where $\boldsymbol{\phi}$ denotes the fixed feature-space mapping, and b the bias parameter. Weight vectors $\boldsymbol{w}$, for which $y^{(i)}(\boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}^{(i)}) + b)$ equals to 1 or -1 lie at the maximum margin hyperplanes and are *support vectors*. SVM separates the

classes so that one class has a positive value for $\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}^{(i)}) + b$ and negative for the other (Bishop, 2006; Goodfellow et al., 2016).

Since the perfect linear separability is not required for the classes in SVM classification, some of the observations are let to be misclassified on the incorrect side of the decision boundary. Therefore, slack variables $\xi^{(i)} \geq 0$ are introduced. Equation (2.13) updates the the optimisation problem and the constraints,

$$\begin{aligned} &\min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|^2 + \gamma \sum_{i=1}^{N} \xi^{(i)} \\ &\text{subject to } y^{(i)}(\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}^{(i)}) + b) \geq 1 - \xi^{(i)}, \quad i = 1, \ldots, N \\ &\qquad\qquad \xi^{(i)} \geq 0, \qquad\qquad\qquad\qquad\quad i = 1, \ldots, N, \end{aligned} \tag{2.13}$$

where $\sum_{i=1}^{N} \xi^{(i)}$ sets an upper bound for the number of misclassified data points, and thus, $\gamma > 0$ is a constant controlling the split between the margin and the slack variable penalty (Cortes and Vapnik, 1995; Hastie et al., 2001; Bishop, 2006). If $\xi^{(i)} = 0$, the data point i has a correct classification as it lies at the margin or on its correct side. $0 < \xi^{(i)} \leq 1$ means also a correct classification, but the data point lies inside the margin but on the correct side of the decision boundary. A data point with $\xi^{(i)} > 1$ is misclassified since it lies on the incorrect side of the decision boundary. Binary SVM classification with slack variables is shown in Figure 2.5.

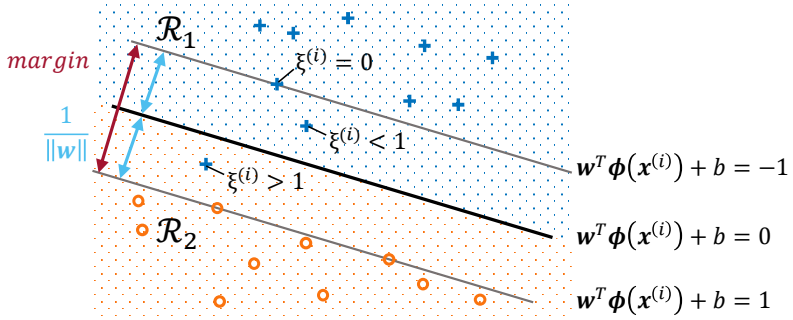


Figure 2.5: The black decision boundary divides the space into regions $\mathcal{R}_1$ and $\mathcal{R}_2$, leaving a margin between the classes in binary SVM classification. Figure following Hastie et al. (2001) and Murphy (2012).

2.3.10 k-nearest neighbours

In k-nearest neighbours classification (k-NN), a data instance is classified to the same class as the majority of its k closest neighbours, and $k = 1, \ldots, \mathrm{N}$ (Bishop, 2006; Hastie et al., 2001; Duda et al., 2001; Mitchell, 1997). If an equal number of neighbours belongs to different classes, the class can be selected, for example, randomly between them. The selected distance measure, such as Euclidean, Mahalanobis, or Manhattan distance, may affect the classification result (Duda et al., 2001). Since k-NN is a non-parametric algorithm, the underlying data are allowed to have any distribution (Goodfellow et al., 2016).

2.4 Evaluating classification results

Measures evaluate and enable to compare the performance of distinct classifiers and the performance of the same classifier with any changes in parameters, features, or other factors (Marsland, 2015). This section provides background for achieving the second research goal of this work by introducing performance measures and by assessing their usability in classification.

2.4.1 Performance measures

Many of the classifiers, presented in Section 2.3, do not provide a predicted label but probabilities in the interval of 0.00–1.00 of belonging to classes (Fawcett, 2006). This probability has to exceed a predefined threshold so that a data point is assigned with a corresponding label. Thereby, the choice of the threshold affects the labelling, and thus, the results. However, the correct threshold varies application-specifically, and selecting the correct one is not straightforward (Saito and Rehmsmeier, 2015). Therefore, single-threshold and threshold-free measures are presented next. For a detailed explanation of the measures, the reader is advised to refer to Sokolova and Lapalme (2009) and Saito and Rehmsmeier (2015).

Confusion matrix is a simple matrix of classification results, and it forms a foundation for classification evaluation. Confusion matrix, presented in Table 2.1, has a size of 2×2 in binary classification. The four sections in the confusion matrix represent how a data point can be classified.

- **True positive** (TP) means a data point, which belongs to class C_1 and is classified to belong to C_1.

Table 2.1: Confusion matrix used in binary classification.

		Predicted class	
		Positive (C_1)	Negative (C_2)
True class	Positive (C_1)	True positive (TP)	False negative (FN)
	Negative (C_2)	False positive (FP)	True negative (TN)

- **False negative** (FN) means a data point, which belongs to class C_1 but is classified not to belong to C_1.
- **False positive** (FP) means a data point, which does not belong to class C_1 but is classified to belong to C_1.
- **True negative** (TN) means a data point, which does not belong to class C_1 and is classified not to belong to C_1.

Single-threshold measures

The following measures require a threshold for the probability of belonging to a class to assess the classification performance. Altering the threshold changes also the number of the four outcomes (TP, FN, FP, TN), and accordingly, the following performance measures (Van Trees, 1968).

Accuracy is the rate of classifying the data instances into the correct classes as defined in Equation (2.14).

$$\text{Accuracy} = \frac{TP + TN}{TP + FN + FP + TN} \tag{2.14}$$

Precision or positive predictive value (PPV) is the rate of data instances with a positive classification, for which the classification is correct as defined in Equation (2.15). In the NICU context, precision means the rate of patients with a complication diagnosis who are truly unwell. A low precision implies that more patients are suspected to have a complication than have that in reality, which means playing it safe in the practical sense.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2.15}$$

Sensitivity, recall, or true positive rate (TPR) is the rate of data instances belonging to the positive class which are classified correctly as defined in

Equation (2.16). Thus, sensitivity measures the rate of identifying the unwell patients from all unwell patients, which is vital from the medical point of view. Not identifying an unwell patient can have critical consequences, and therefore, false positives are much more acceptable than false negatives (Rollins et al., 2015).

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{2.16}$$

Specificity or true negative rate (TNR) is the rate of data instances belonging to the negative class which are classified correctly as defined in Equation (2.17). Thereby, specificity measures the rate of truly healthy patients, which have been diagnosed as healthy.

$$\text{Specificity} = \frac{TN}{FP + TN} \tag{2.17}$$

False positive rate (FPR) is the rate of data instances belonging to the negative class which are classified incorrectly as defined in Equation (2.18). At NICUs, FPR is the rate of healthy patients, which are diagnosed as sick.

$$\text{False positive rate} = 1 - \text{specifity} = \frac{FP}{FP + TN} \tag{2.18}$$

$\mathbf{F_1}$ score, F-score, or F-measure, defined in Equation (2.19), is the harmonic mean of precision and sensitivity.

$$\text{F}_1 \text{ score} = \frac{2 \cdot \text{precision} \cdot \text{sensitivity}}{\text{precision} + \text{sensitivity}} \tag{2.19}$$

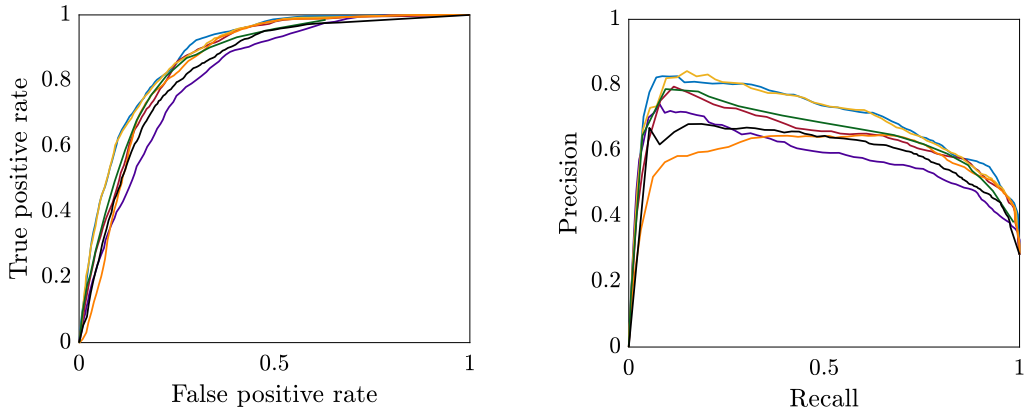
Threshold-free measures

The following measures merge single-threshold measures so that all possible thresholds in the range of 0.00–1.00 are taken into account.

Receiver operating characteristics (ROC), example in Figure 2.6(a), visualise the results of a binary classification task (Hanley and McNeil, 1982; Fawcett, 2006). The false positive rates (FPRs) for all thresholds lie on the x-axis, and they are plotted against the true positive rates (TPRs) for all thresholds on the y-axis (Van Trees, 1968; Fawcett, 2006; Davis and Goadrich, 2006; Saito and Rehmsmeier, 2015). The ROC curve of a perfect classification passes from (0,0) through (0,1) to (1,1). Random guessing produces a

diagonal ROC curve, from left bottom corner to right top corner. Therefore, only classifiers in the upper left triangle outperform random guessing.

Area under the ROC curve (abbreviated as AUROC in this work) is a single value between 0 and 1, which makes comparing ROC curves of distinct classifiers more convenient (Hanley and McNeil, 1982; Saito and Rehmsmeier, 2015; Fawcett, 2006). If AUROC is 1, the two groups have been identified perfectly and they are totally distinct whereas an AUROC value of 0.5 implicates random guessing and the groups have not been identified at all (Fawcett, 2006; Swets, 1988; Griffin and Moorman, 2001). Thus, all classifiers should have an AUROC higher than 0.5. Noteworthy, AUROC quantifies only the area, not the shape of the curve, and two distinct ROC curves can have the same AUROC.



(a) ROC curves (b) PR curves

Figure 2.6: Results for seven classifiers in terms of ROC and PR curves.

Precision-recall (PR) curve is another classification performance measure, illustrated in Figure 2.6(b). The values of recall for all thresholds lie on the x-axis, and they are plotted against the values of precision for all thresholds on the y-axis (Davis and Goadrich, 2006; Saito and Rehmsmeier, 2015). The perfect classification lies in the upper right corner.

Area under the PR curve (abbreviated as AUPR in this work) quantifies the PR curve into a single value between 0 and 1, making comparisons between classifiers more convenient.

2.4.2 Applicability of measures

Since the classification results can be assessed with many measures, the choice of the measure depends on what is wanted to be measured. Thereby, the choice is also a matter of opinion. The different measures emphasise different aspects as described in Section 2.4.1. However, only appropriate evaluation criteria provide justified results that answer to research questions. Therefore, it is essential to be aware of the capabilities and limitations of different measures (Fawcett, 2006). For example in complication predictions, the interest is often in identifying the sick patients among all patients, which means classifying the positive class correctly. Therefore, the most suitable measures are required to concentrate on evaluating that.

The suitability of measures for assessing the classification performance depends partly also on the underlying data. Data imbalance (see Section 2.5) means that the ratio of the positive and negative classes is not equal, forming majority and minority classes. This disproportion affects the choice of the appropriate measure. For example, accuracy is not the optimal evaluation criterion for imbalanced datasets if the task is to identify the minority class representatives (Libbrecht and Noble, 2015; Marsland, 2015; Rollins et al., 2015; Rokach, 2010). If the sick patients were the minority class and the healthy patients the majority class, classifying all patients to the class of the healthy would result in a high accuracy even though none of the sick patients was identified and classified correctly. Therefore, the use of other measures is required, and Marsland (2015) expresses that either the pair of precision and recall or the pair of specificity and sensitivity provides more information than accuracy alone.

In medical data, the class of sick subjects is often the minority class. Accordingly, class imbalance has to be considered in medicine since misclassifying sick patients as healthy can be vital for them if they do not receive medical care on time (Weiss and Provost, 2001). Therefore, it is important to identify all sick patients, which implies that a high sensitivity is appreciated. Even though misclassifying healthy patients as sick is not harmful for them, it is waste of resources to consider and treat them as risk patients in vain. Therefore, it is essential to classify only the sick patients as sick, which implicates that a high precision is valuable. In accordance, precision and sensitivity are more appropriate evaluation criteria than accuracy (Sun et al., 2009).

The single-threshold measure F_1 score is a derivative of precision and sensitivity, and using F_1 score is supported from the data imbalance point of view (Marsland, 2015; Sun et al., 2009). F_1 score evaluates the ability of

the classifier to truly identify the data points of the underrepresented class, and it does not provide overly optimistic results either as accuracy and some other measures do. The same applies to the threshold-free AUPR that is another derivative of precision and sensitivity (recall).

Saito and Rehmsmeier (2015) researched the performance of ROC and PR curves on balanced and imbalanced datasets, concluding PR curves result in more informative and intuitive plots if data imbalance is present. However, this is debatable since Fawcett (2006) encourages the use of ROC curves over PR curves due to their resistance to changes in class balance. For all that, according to Saito and Rehmsmeier (2015), ROC curves are used more frequently in the studies, and the statement is supported by the findings in Section 2.6. Only a few researchers have reported other measures: Rollins et al. (2015) have reported F_1 score and Desautels et al. (2016) AUPR.

2.5 Challenges in clinical data

Hogan and Wagner (1997) describe the data quality with two measures: *correctness* is the proportion of truly correct data observations to incorrect data observations, and *completeness* is the proportion of recorded observations to all recordable observations. Both correctness and completeness are important factors regarding the performance of machine learning algorithms. Generally speaking, medical data and physiological parameter recordings are seldom totally correct or complete. The data are sparse and noisy, the sampling is irregular, and the data samples are plagued by human error (Ghassemi et al., 2015; Marlin et al., 2012). Additionally, some values may be out of range, and there can be gaps in the time series (Salcedo-Bernal et al., 2016). Additionally, some of the missing values are caused by probe dropouts such as malfunctions or removals of the measuring equipment (Stanculescu et al., 2014a). Consequently, all these decrease the correctness and completeness of the data.

Missing values

Missing values mean gaps in the data or the sparsity of the data. They increase the level of incompleteness of the data, which is characteristic for many real-world data sets (Donders et al., 2006; Kotsiantis et al., 2006). Sometimes preprocessing the data produces missing values. For example, Lehman et al. (2008) replaced measurement values out of range by missing values, but filled them later by interpolation.

While statistical methods function well with data that contain noise and missing values, predictive methods often fail with such data. Therefore, many techniques have been developed to deal with missing values. Saar-Tsechansky and Provost (2007) suggest four alternative approaches to handle missing values as follows.

1. The whole data instance $(\boldsymbol{x}^{(i)}, y^{(i)})$ with a missing value is discarded.
2. The whole feature $\boldsymbol{x}_j$ with a missing value is discarded.
3. The missing value $\boldsymbol{x}_j^{(i)}$ is acquired.
4. The missing value $\boldsymbol{x}_j^{(i)}$ is estimated by
   (a) replacing it with the mean or mode of the feature j,
   (b) replacing it with an arbitrary unique value, or
   (c) calculating it from the distribution of the feature j.

To extend suggestion 4(a), multiple extrapolation methods have been proposed to fill the missing values by, for example, with the mean of the whole data or the mean of the adjacent values (Meyfroidt et al., 2009). Furthermore, a simple last-observation-carry-forward (LOCF) method has been applied (Desautels et al., 2016; Overall et al., 2009; Mani et al., 2014). In LOCF, missing values are replaced with the previous known value.

In addition, more sophisticated and complex methods have been proposed. The generative probabilistic models, such as autoregressive hidden Markov models, are appropriate for estimating missing values as they utilise marginalisation (Stanculescu et al., 2014b). Marginalisation means drawing probabilities for unknown values from the known values, and the direct dependencies between all values are taken into account. However, these models require the proportion of missing values to be relatively small. Further, generalised linear mixed models can be applied on sparse data as they function despite the missing values (Overall et al., 2009). Still, the use of simpler models is advised due to their better performance.

Irregular sampling

Irregular sampling means that the time intervals between samples do not stay constant. This irregularity causes many modelling methods to fail, which can, however, be tackled by making assumptions about the functional form of the data (Ghassemi et al., 2015). Of course, making assumptions introduces new bias to the model.

A technique to tackle the varying sampling frequency is piecewise aggregate approximation (PAA), in which the time series are cut into time frames of equal length (Keogh et al., 2001). Then, the values in each time frame are

averaged for each time frame. In cases where no values exist in the frame, the same value is selected as in the previous or the following frame (Salcedo-Bernal et al., 2016). Marlin et al. (2012) applied PAA with one-hour-long intervals and mean filtering but they also pointed out the issue of potential information loss. Despite not calling it PAA, Lehman et al. (2008) used a similar approach with one-minute-long time intervals where medians of the samples were calculated for each minute, and Lehman et al. (2015) had the same interval length but calculated averages.

A time series can also be translated into a string of symbols to avoid the challenges, caused by irregularly sampled time series. One symbolic method is symbolic aggregate approximation (Lin et al., 2007). First, this method uses PAA to split the time series into frames of equal length, each of which is assigned with the mean value of that frame. Then, the mean values are discretised by setting *breakpoints* B for their values, which are used to assign each time frame with a symbol such as an alphabet. For example, two breakpoints $B = \{\beta_1, \beta_2\}$, $\beta_1 < \beta_2$, are set for a PAA representation. Values below β_1 are given an A, the values between β_1 and β_2 a B, and the values above β_2 a C. The breakpoints B are advised to be derived from the Gaussian distribution (Lin et al., 2007).

Ghassemi et al. (2015) proposed a time series modelling method to make predictions from clinical data. Their method uses multiple irregularly sampled time series along with their between and within correlations. This multivariate method introduces a new latent space and uses the multi-task GP models, outperforming univariate time series methods.

Imbalanced data

Class imbalance means the disproportional occurrence of class representatives in the data, leading to majority and minority classes. Imbalanced data are problematic especially in binary classification if the class of interest is the minority class (Cerqueira et al., 2014). As the model has not been trained with a sufficient number of minority class representatives, many classifiers fail in classifying the minority class correctly (Weiss and Provost, 2001; Marsland, 2015). In these cases, the classifier does not necessarily learn – or is even not trained with – all possible variations of the minority class representatives.

The class imbalance can be managed with resampling. In *oversampling*, the minority class samples are copied at random until their number has increased close to the number of the majority class samples, and in the opposite case, in *undersampling*, the majority class samples are removed at random until their number has decreased close to the number of the minority class sam-

ples (Japkowicz and Stephen, 2002; Estabrooks et al., 2004). There is no unambiguous solution which resampling technique to use since their performance depends on the underlying data (Estabrooks et al., 2004). Selecting either often improves the result compared to using the imbalanced data. Nevertheless, Japkowicz and Stephen (2002) conclude that oversampling outperforms undersampling.

Improving the data quality with expert knowledge

Besides knowledge of data science, substance knowledge is required to select the most important features for the machine learning algorithms at the data preprocessing phase (Cerqueira et al., 2014). Adding expert or background knowledge in machine learning has also been considered so that the machine learning models would not depend only on the input data but also on the clinical expertise (Lucas, 2004; Bellazzi and Zupan, 2008). Holzinger (2016) discusses the possibility to create interactive machine learning algorithms, where an expert is involved in the actual learning phase of machine learning algorithms, in addition to the preprocessing phase. Nonetheless, this *"human-in-the-loop"* approach lacks quantitative research on its performance and suitability in health care and medicine.

Comparability

Salcedo-Bernal et al. (2016) point out the difficulty to compare the results of different research papers in the clinical field since the applied data and parameters vary from paper to paper, thus making it hard to conclude which model gives the most accurate predictions. Recently, the open MIMIC II (Multiparameter Intelligent Monitoring in Intensive Care) database (Saeed et al., 2011), available on Physionet (Goldberger et al., 2000), has been used by several researchers, such as Salcedo-Bernal et al. (2016), Lehman et al. (2015), Ghassemi et al. (2015), and Calvert et al. (2016).

2.6 Previous work

The results of a comprehensive literature review to the previous work of machine learning applications in neonatology, at ICUs, and in health care in general are provided in this section. The common denominator for all the studies presented here is the data-based approach to predict mortality or a medical complication.

2.6.1 Mortality predictions

Neonatal mortality has been studied with GP and SVM classifiers using measurements from five physiological time series, GA, BW, and SNAP-II and SNAPPE-II scores (Rinta-Koski et al., 2017a, 2018). Rinta-Koski et al. (2017a) studied the impact of feature selection of 24-hour-long time series using GP with the kernel presented in Equation (2.5d). They showed the highest AUROC to be 0.94 with many different feature combinations. Selecting only the time series features decreased the result slightly to 0.88. Rinta-Koski et al. (2018) extended the research to cover also time series of other lengths, and they included SVM classifier and GP classifiers with other kernels in Equations (2.5a), (2.5b), and (2.5c) in the study. As a result, GP classifiers outperformed SVM, and the optimal length of the monitoring time was 48 hours from the birth. Similarly to Rinta-Koski et al. (2017a), different feature combinations were tested. If only time series features were used, they showed an AUROC of 0.926. That remains lower than 0.947 or 0.949, which were achieved by combining time series features with GA and BW, or GA, BW, and the medical scores SNAP-II and SNAPPE-II, respectively.

Neonatal mortality was also studied by Cerqueira et al. (2014) who, first, applied statistical analyses and medical experts to select the preferred features for the model. Majority of the features were single values, such as binary indicators of the presence of a certain complication or the occurrence of a certain treatment. Then, they applied SVM and artificial neural networks to predict the death of patients and achieved AUROCs of 0.83 and 0.84, respectively.

Salcedo-Bernal et al. (2016) predicted the in-hospital mortality at an ICU using multivariate time series of heart rate, respiratory rate, and SpO_2. They compared LR, neural networks, k-NN, and DT classifiers and received accuracies of 0.68, 0.75, 0.65, and 0.74, respectively. Optimising the parameters of the models did not improve the results in logistic regression and neural networks.

Lehman et al. (2015) utilised time series of heart rate and blood pressure as well as the medical scores APACHE III, APACHE IV, and SAPS to predict in-hospital mortality of ICU patients with a switching vector autoregressive framework (Murphy, 1998; Nemati et al., 2012). The highest results are received by selecting blood pressure along with one of the scores at a time as features. Blood pressure alone results in an AUROC of 0.70, while SAPS increases it to 0.77 (SAPS alone 0.65), APACHE III to 0.84 (APACHE III alone 0.80), and APACHE IV to 0.85 (APACHE IV alone 0.82).

Ramon et al. (2007) predicted the mortality at an ICU from a large dataset which contain, among others, the patient basic information, physiological parameter measurements, and medication details. The classification performance was measured with AUROC, which was 0.79 for DT, 0.82 for first order RF, 0.88 for NB, and 0.86 for tree augmented NB. They also predicted a number of different complications, concluding RF classifiers always outperform DTs.

In addition to supervised learning, unsupervised machine learning can be applied to identify the patients in danger to die. Marlin et al. (2012) studied mortality in ICU environment using clustering and mainly physiological parameter measurements. They resulted in an AUROC of approximately 0.85–0.90. The performance was improved if the length of parameter monitoring was prolonged.

In addition to the aforementioned mortality research, many other studies have been conducted. The reader is advised to refer to Medlock et al. (2011) who have made a comprehensive review of existing studies on the prediction models of mortality, focusing solely on VLBW and VLGA infants. The number of the identified studies is 41 and the majority of them, 35 to be accurate, have used logistic regression to predict mortality.

2.6.2 Morbidity predictions

Besides predicting morbidities in general, predicting neonatal morbidities has been in the interest of a decent amount of research. These morbidities include, but are not limited to, BPD, NEC, ROP, and sepsis. In the early 2000s, the focus was on identifying the most relevant features that are either capable of detecting or predicting a certain morbidity. The features were usually selected among the patient basic information, such as GA or the presence of a certain complication. In the recent years, numerous machine learning approaches have been proposed to predict morbidities from various types of data, including monitored sensor values or laboratory test results.

Saria et al. (2010) predicted BPD, intraventricular haemorrhage, NEC, ROP, and death from GA, BW, and the physiological parameters of heart rate, respiratory rate, and oxygen saturation with Bayesian modelling. Predicting any of the aforementioned morbidities or death, they achieved an AUROC of 0.92. The medical scores Apgar, CRIB, SNAP-II, and SNAPPE-II alone resulted in 0.70, 0.85, 0.83, and 0.88 respectively. They also compared the performance of their method and the medical scores for infections, such as NEC, sepsis, and urinary tract infection, and cardiopulmonary complications,

such as BPD, resulting in AUROCs of 0.97 and 0.98 compared to 0.74 and 0.72, 0.90 and 0.91, 0.84 and 0.86, and 0.91 and 0.93 for Apgar, CRIB, SNAP-II, and SNAPPE-II scores, respectively. Their final observation was that including all features in the model shows a higher AUROC of 0.91 compared to including only GA and BW (AUROC 0.85) or only physiological parameters (AUROC 0.85).

Rinta-Koski et al. (2017b) predicted BPD, NEC, and ROP with a GP classifier using the mean and standard deviation of five physiological time series as well as GA, BW, and SNAP-II and SNAPPE-II scores. They also studied the effect of feature selection on the results. They were able to achieve an AUROC of 0.87 for BPD. Even though AUROCs were 0.74 and 0.84 for NEC and ROP, respectively, predicting them was not successful as the sensitivities were close to zero.

Bronchopulmonary dysplasia

The previous research has predicted neonatal BPD from a variety of features that have mainly been patient basic information or indicators of the presence of a certain complication or treatment. However, the use of physiological time series as features is limited. In contrary, many classifiers have been applied to study which classifier is the most suitable to predict BPD. Nevertheless, no general consensus exists for the optimal classifier even though majority of the research has focused on logistic regression and some papers apply neural networks or SVMs (Ochab and Wajs, 2016).

Wajs et al. (2006) used BW, a binary variable of the presence of respiratory support, alveolar-arterial ratio, a binary variable of the presence of patent ductus arteriosus, SpO_2, and heart rate as features in logistic regression to predict neonatal BPD. They received an AUROC of 0.942.

Furthermore, Wajs et al. (2007) examined all possible combinations of 14 features. The optimal features consisted of BW, a binary variable of the presence of patent ductus arteriosus, surfactant administration, a binary variable of the presence of respiratory support, ratio of time when SpO_2 is below 85 %, mean heart rate, and the ratio of mean SpO_2 during the first week to mean SpO_2 during the first day. LR and RBF neural network were applied on these features, resulting in AUROCs of 0.91 and 0.95, respectively.

Ochab and Wajs (2014b) compared various combinations of the same 14 features, and predicted BPD with both SVM and LR, both implemented in Matlab. Despite the feature combination, LR outperformed SVMs in terms of accuracy and sensitivity. Interestingly, the implementation environment affected the results since Ochab and Wajs (2014a) repeated the experiments

with the LIBSVM library by Chang and Lin (2011). This time, SVM was able to achieve a better accuracy and sensitivity for certain feature combinations than LR, outperforming usually also the Matlab implementation of SVM. Furthermore, Ochab and Wajs (2016) studied the feature selection for the same task using LIBSVM and LR classifiers. They drew a conclusion that LR provides a higher accuracy when the number of features is less than seven, whereas LIBSVM functions better when more than seven features are included. Finally, Wajs et al. (2018) predicted BPD from the same features using NB classifier which was outperformed by either LR or SVM, depending on the performance measure.

Multiple studies have applied logistic regression and other statistical methods to identify the features associated with neonatal BPD. These studies, however, have seldom used physiological time series as features but rather static values. Bhering et al. (2007) used four variables, which include GA and the presence of mechanical ventilation, and received an AUROC of 0.935. Cunha et al. (2005) found eight features, such as BW, GA, and presence of patent ductus arteriosus, to be associated with a developing BPD. Romagnoli et al. (1998) used similar features and showed an AUROC of 0.960 for infants at the age of 72 hours. Kim et al. (2005) used GA, BW, Apgar score and five other features to predict BPD, resulting in AUROCs of 0.90, 0.91, and 0.94 at the ages of four, seven, and ten days, respectively.

Using not only LR but also tree models, Ambalavanan et al. (2008) were able to associate lower BW, higher oxygen concentration, male gender, additional surfactant doses, higher oxygenation index, and outborn status with a higher risk of BPD and death. However, they acknowledged that more validation is required due to the limited number of patients in the study.

Laughon et al. (2011) compared the effect of feature selection on the classification performance and discovered that six optimal features are GA, BW, race and ethnicity, sex, respiratory support, fraction of inspired oxygen. These features resulted in AUROCs of 0.793 and 0.854 at the first and the 28^{th} day of life, respectively.

Necrotising enterocolitis

The research to predict NEC with machine learning algorithms is limited, but a few applications are able to distinguish patients with NEC from those without as well as the required treatment, surgical or medical. In fact, no biological indicator of NEC is currently used in practise due to their low predictive power (Sylvester et al., 2014). Therefore, the current practise diagnoses NEC clinically instead of diagnostic tests (Ji et al., 2014).

Since NEC and sepsis have similar pathophysiologic features, Stone et al. (2013) used a successful technique to predict sepsis, namely heart rate characteristics (HRC) index, and extended it to NEC predictions felicitously. Stone et al. (2013) observed that the baseline of HRC index rises for the patients requiring a surgical intervention, and the rise appears for 1–3 days prior to the NEC diagnosis. Additionally, a significant increase in the HRC index is detected for 16 hours prior to the diagnosis of surgical NEC and for 6 hours prior to that of medical NEC.

Sylvester et al. (2014) investigated if it was possible to predict the treatment type, surgical or medical, beforehand from clinical parameters or biomarkers. Using 27 clinical input features for LDA, the algorithm distinguishes the two types with an AUROC of 0.817, whereas three specific biomarkers show a higher AUROC of 0.856. Were the clinical parameters and biomarkers used together, the treatment groups are distinguished perfectly. The statistically most significant parameters in the analysis were male gender and BW.

Ji et al. (2014) used the same 27 clinical parameters as Sylvester et al. (2014) and applied LDA to predict the level of risk for NEC using three categories: low, intermediate, and high. This prediction received an AUROC of 0.85.

Retinopathy of prematurity

Approaches to predict ROP mainly use retinal images or basic patient information as input. In addition, statistical research has been conducted to select the most predictive and revealing features of ROP. Unfortunately, continuous physiological measurements as the input data lack research.

Bolón-Canedo et al. (2015b) performed a comprehensive study on the usability of machine learning in predicting ROP from retinal images. First, they compared six feature selection algorithms, all of which produced similar results. Second, they compared these features to features which were selected by a group of experts. Similarities in these two feature sets were remarkable. Third, they performed binary classification of the patients using DT, NB, k-NN, SVM, and compared the results to the classification made by the experts. The algorithms achieved at least as good results as the experts. The lowest classification errors (less than 0.11) were achieved by NB and SVM. However, developing a golden standard for the process of predicting ROP is difficult, which this study did not reach either.

Ataer-Cansizoglu et al. (2015) used Gaussian mixture model to extract features from retinal images and used SVM to classify patients. They received an accuracy of 0.95 which approximately equals to the accuracy of classification made by experts. Also, Bolón-Canedo et al. (2015a) used Gaussian

mixture model to extract features from retinal images. They combined those features with traditional statistical features, and used DT, NB, SVM, and RF for classification. SVM achieved the highest accuracy of 0.911 when all features were included in the model, whereas the features from Gaussian mixture model alone had an accuracy of 0.905. Furthermore, various image analysis approaches have been proposed to detect the patients at risk of ROP (Wittenberg et al., 2012).

Rollins et al. (2015) proposed a discrete conditional phase-type model that functions with class imbalance. The model requires a classifiers as a component in the model. SVM component was shown to outperform DT and RF components when ROP was predicted for VLGA and VLBW infants. They achieved an F_1 score of 0.738.

Rather than machine learning, Löfqvist et al. (2006) used logistic regression to select the features for ROP predictions. The optimal features are postnatal weight gain, insulin-like growth factor level, and insulin-like growth factor binding protein 3 level, all measured on a weekly basis. Further, Wu et al. (2012) simplified the algorithm to include only the weekly weight gain and were still able to predict ROP.

Binenbaum et al. (2011) used LR to observe that BW, GA, and postnatal weight gain provide the highest prediction performance for the risk of ROP. Darlow et al. (2005) performed statistical analyses and LR to define the most significant variables for ROP predictions. They were able to associate a low GA, a low BW among other preterm infants with the same GA, and the male gender with an increased risk of developing ROP.

Sepsis

Sepsis causes sudden clinical deterioration of neonates and is a major reason behind neonatal morbidities and mortality (Griffin and Moorman, 2001; Griffin et al., 2003). Therefore, detecting sepsis as early as possible is important so that more aggressive and targeted treatment can be started on time (Desautels et al., 2016). However, diagnosing sepsis from clinical signs and laboratory tests beforehand has been proven to be difficult (Escobar, 1999; Griffin and Moorman, 2001). To make more reliable diagnoses for sepsis, several studies have been conducted on the heart rate characteristics (HRC) and their abnormalities, such as reduced variability and transient decelerations (Griffin and Moorman, 2001; Kovatchev et al., 2003; Griffin et al., 2003, 2004, 2005; Moorman et al., 2006). Furthermore, the most appropriate features for a predictive model have been identified from physiological time series and laboratory test results. Machine learning and statistical mod-

els have been used to predict neonatal sepsis and sepsis in general, and the predictive results have been compared to predictions of medical scores (Stanculescu et al., 2014a,b; Calvert et al., 2016; Desautels et al., 2016; Mani et al., 2014; Wang et al., 2013).

Griffin and Moorman (2001) examined the relationship between HRC, such as its statistical moments and percentiles, and the risk of neonatal sepsis using multivariate logistic regression. The explanatory power of scores SNAP and NTISS on the occurrence of sepsis was also studied. They observed that HRC are abnormal for 24 hours prior to the clinical suspicion of sepsis with an AUROC of 0.90 – especially skewness and percentiles revealed the patients at risk. As sepsis deteriorates the physiological parameters of the patients, there is a rise in both SNAP and NTISS scores before the sepsis suspicion. SNAP is affected significantly more than NTISS. What is more, the infants with sepsis tend to have a BW of approximately 200 g less and a GA of approximately two weeks less than healthy infants.

To have more evidence for the results, Griffin et al. (2003) showed that there is a significant connection (AUROC 0.75) between HRC index and neonatal sepsis and other sepsis-like illnesses. Further, Griffin et al. (2004) used multivariate logistic regression to show that HRC index had an association with death up to seven days in advance (the highest AUROC of 0.74) and the cumulative HRC index was associated with in-hospital mortality (AUROC of 0.83). In addition, using HRC index together with BW, GA, and postnatal age in a multivariate logistic regression was shown to increase the predictive power of sepsis: AUROC increased from 0.75 to 0.77 while the additional features result in 0.67 alone Griffin et al. (2003). Also, the AUROC of death predictions increased from 0.74 to 0.85 with the additional features, which resulted in 0.70 alone (Griffin et al., 2004). However, combining the parameters with the cumulative HRC index for predicting in-hospital mortality reduced the AUROC from 0.83 to 0.79. The parameters alone had an AUROC of 0.76. Additionally, Griffin et al. (2005) combined HRC index with certain laboratory test results, which improved the AUROC of neonatal sepsis predictions from 0.73 to 0.82. Laboratory tests alone resulted to an AUROC of 0.75.

Kovatchev et al. (2003) studied sample asymmetries of heart rate variability in order to detect neonatal sepsis or systemic inflammatory response syndrome. Their results showed the sample asymmetries grow before the diagnosis and treatment of the complication.

Calvert et al. (2016) used machine learning algorithms to predict sepsis of ICU patients from the following physiological time series: systolic blood pres-

sure, pulse pressure, heart rate, temperature, respiration rate, white blood cell count, pH, blood oxygen saturation, and age. They were able to reach an AUROC of 0.92 at three hours before a systemic inflammatory response syndrome period, and an AUROC of 0.83 at less than three hours before that period.

Desautels et al. (2016) predicted sepsis of ICU patients from eight physiological time series: systolic blood pressure, pulse pressure, heart rate, respiration rate, temperature, peripheral capillary oxygen saturation, age, and Glasgow coma score. They applied binary classification on the patients, receiving an AUROC of 0.88 and an AUPR of 0.60 at the sepsis onset. This result outperformed the performance of multiple other medical scores, which were 0.80 and 0.33 for MEWS, 0.70 and 0.23 for SAPS II, 0.61 and 0.16 for systemic inflammatory response syndrome by Bone et al. (1992), 0.73 and 0.28 for SOFA, and 0.77 and 0.28 for quickSOFA in terms of AUROC and AUPR, respectively. Differences between the scores were explained by the different input data requirements. The algorithm functioned well also with sparse data where up to 60 % of the data were missing.

Mani et al. (2014) compared many classifiers to predict neonatal sepsis. The classifiers were tested on data both including and excluding the culture negative sepsis patients, receiving the following AUROCs for the two datasets: RF (0.57, 0.65), classification and regression trees (0.65, 0.77), SVM (0.61, 0.68), k-NN (0.54, 0.62), LR (0.61, 0.61), lazy Bayesian rules (0.62, 0.58), NB (0.64, 0.78), and tree augmented naïve Bayes (0.59, 0.53). All classifiers had higher specificity than predictions made by physicians, and almost all classifiers also exceeded the sensitivity of the experts' predictions.

Wang et al. (2013) used embedded methods in feature selection to identify the optimal biomarkers from ten alternatives, such as white blood cell count and haemoglobin count, to predict neonatal sepsis. They applied canonical correlation analysis to identify the optimal features and sparse support vector machine classifier to test their predictive power. The highest accuracy of 0.875 was achieved with five features.

Also unsupervised learning has been applied to predict sepsis. Based on the factorial switching linear dynamical system by Quinn et al. (2009), Stanculescu et al. (2014a) developed a deep learning styled hierarchical switching linear dynamical system which takes the complex interactions of a dynamic system into account. They used time series of heart rate and SpO_2 measurements to predict sepsis of VLBW infants since lowered heart rate and SpO_2 are often indicators of sepsis. They resulted in an AUROC of 0.69.

To have a comparison for the results of Stanculescu et al. (2014a), Stanculescu

et al. (2014b) performed the same analysis on the same time series to predict sepsis, but this time using a few additional measurements and autoregressive hidden Markov models. They received an AUROC of 0.74 when all data were used and an AUROC of 0.72 without any missing values.

2.7 Background conclusions

The Section 2.1 discusses the need for research in the field of neonatology where the preterm infant patients are prone to critical morbidities and mortality. Their lives may be saved if the occurrence of the morbidities or death can be predicted in advance as it enables more time for the medical doctors to treat them. Therefore, conducting this study is important, and the other sections in Chapter 2 present topics that are relevant in order to achieve the three research objectives of this study.

The first research objective is to identify the most suitable classifiers for predicting neonatal mortality and morbidities, and thus, 12 classifiers are presented in Section 2.3. However, Meyfroidt et al. (2009) concluded that no classifier is more suitable than any other, and Ochab and Wajs (2016) stated that no model has been generally accepted for predicting BPD. Currently, LR has been applied the most widely in this field as Section 2.6 and Medlock et al. (2011) reveal, followed by SVM and DT classifiers. QDA is the only algorithm of the presented classifiers that lacks research. Based on the previous studies, it is difficult to compare the performance of different classifiers and name the most suitable classifier due to the dissimilar underlying data (Salcedo-Bernal et al., 2016). Therefore, this study compares the performance of the 12 classifiers on the same data.

What comes to the predictability of complications, the literature review in Section 2.6 shows successful proposals in predicting mortality and BPD from varying types of patient data. The mortality predictions have achieved AUROCs of 0.88 with NB, 0.83 with SVM, 0.82 with RF, 0.79 and 0.74 with DT, 0.68 with LR, and 0.65 with k-NN. Previous BPD predictions have, on the other hand, reached high AUROCs with logistic regression that are 0.94 and 0.91. There is less research on predicting NEC and ROP, and the results are more modest. No indicator has been proven to be powerful enough to predict NEC, and the research is focused on predicting the type of required treatment (Sylvester et al., 2014). Many ROP studies, such as Ataer-Cansizoglu et al. (2015) and Wittenberg et al. (2012), use retinal images as the input data, raising the question if physiological parameters even reveal the devel-

oping ROP. However, a demonstrated F_1 score of 0.738 for ROP predictions exists.

The second research goal arises from choice of a relevant measure for evaluating the classification algorithms. There are challenges in assessing their performance on imbalanced medical data where the ratio of sick patients to all subjects is low. As discussed in Section 2.5, the data imbalance often causes classifier algorithms to misclassify subjects, which is life-threatening for the sick patients that are classified as healthy (Weiss and Provost, 2001). According to Saito and Rehmsmeier (2015) and the observations in Section 2.6, majority of the previous studies use accuracy and AUROC for performance evaluation. Using them is not advised as they do not solely focus on evaluating the identification of sick patients but they show a good result for identifying healthy patients as well. Instead, using precision, sensitivity, F_1, and AUPR is recommended since they provide more truthful results (see Section 2.4 for reasoning). An example of the optimistic results of AUROC is the result of Desautels et al. (2016): the reported AUROC value is 0.28 units higher than AUPR value for the binary classification results and almost 0.50 units higher for all score-based classifications.

The third objective is to study factors in preprocessing and feature selection. The effect the time series sampling on the classification lacks research in this field, but multiple techniques to handle irregularly sampled time series have been proposed. Also, the impact of the length of patient monitoring time has not been studied adequately since most research uses only static, not temporal features. Rinta-Koski et al. (2017b) concluded that a monitoring time of 72 hours performs slightly better than 24 hours, Rinta-Koski et al. (2018) observed the highest AUROC values at 48 hours, and Marlin et al. (2012) noted that a longer monitoring time improves results.

However, the optimal feature selection has been studied widely since it can improve the classification results (Guyon and Elisseeff, 2003). Using any kind of physiological data or other patient information outperforms the results of the medical scores as, for example, Lehman et al. (2015) and Desautels et al. (2016) have shown. GA and BW are used in most of the studies, and they correlate with the risk of complications (Gomella et al., 2013; Fattore et al., 2015). Combining other features with GA and BW usually improves the classification performance (Saria et al., 2010; Rinta-Koski et al., 2018). In addition, using medical scores or pure time series data alone does not result in the highest performance but combining them with other features improves the results. All in all, adding more features in the model usually improves the performance.

3. Materials and Methods

3.1 Data

The Section 3.1 describes the patient cohort, based on which this study is conducted. In addition, the data quality is evaluated critically.

3.1.1 Data collection and storing system

The neonatal intensive care unit at Children's Hospital, Helsinki University Hospital has been collecting and storing the clinical data of their patients using Clinisoft clinical information management system since 1999. Besides Helsinki University Hospital, the same Clinisoft system is in use and used for research purposes also at the intensive care units at other university hospitals in Finland (Seppänen et al., 2016), at Karolinska University Hospital in Stockholm, Sweden (Honoré, 2017), and at Onze Lieve Vrouwe Gasthuis Teaching Hospital in Amsterdam, the Netherlands (Bosman et al., 1998).

Clinisoft clinical information management system (GE Healthcare, Helsinki, Finland; along with its predecessors) is a brand of electronic health records (EHRs). According to the definition of International Organisation for Standardization (ISO/TR 20514:2005(E), 2005), EHRs are information repositories, accessible only by the authorised users, to store patients' retrospective, concurrent, and prospective health data in a standardised format. However, the practises to store data in EHRs vary case-specifically. Häyrinen et al. (2008) state that EHRs can contain anything only from a few files to comprehensive and longitudinal datasets, whereas Zhao et al. (2017) emphasise that longitudinal data is stored in EHRs. However, the ISO standard does not require the data to be longitudinal (ISO/TR 20514:2005(E), 2005). Despite the amount of data in EHRs, their content is related to patients' hospital or health centre visits, including, for example, measurement values from patient monitoring, laboratory test results, medical diagnoses, medication details, or

clinical notes in unstructured, free text form (Jensen et al., 2012; Häyrinen et al., 2008; Zhao et al., 2017; Meyfroidt et al., 2009).

Constructing and preparing the database for research purposes and the transfer of the database from the hospital's EHR to the university environment has been completed previously as a part of the doctoral dissertation by Rinta-Koski (2018). The database has been implemented in an open source database management system `PostgreSQL` (PostgreSQL Global Development Group, 2019), and the data of interest have been retrieved from the database using SQL queries. Constructing the database has also included pre-cleaning the data as there have been inconsistencies in the registration practices. For instance, the weights have initially been reported either in grams or kilograms, and they have been shifted to the same units.

3.1.2 Data description

The research permission of this study enables to access to the data of VLBW infants which entered the NICU at Children's Hospital, Helsinki University Hospital during 1999–2013. Therefore, the newer data entries are not considered in this study, and the total number of patients is 2059. Due to ethical reasons, the data have been pseudonymised so all identifying factors, such as names and personal identity codes, have been removed from the data. Nevertheless, the data entries of individual patients have been allocated to database-specific identity numbers to keep them connected.

The EHR contains information that has not been collected during the stay at the NICU. It is non-temporal, patient-specific basic information such as birth and gender details, birth weight, gestational age, blood group, Apgar score, and time of entering the NICU. The median GA is more than 28 weeks (accurately 202 days) with a standard deviation of more than 8 weeks (accurately 61 days), and the median BW is 1.105 kg with a standard deviation of 0.287 kg. During the stay at NICU, various types of temporal data are collected from the patient. These data are related to the physiological parameters of patients' diagnoses, medical procedures, laboratory tests, medication, and nutrition.

There are 111 measured and automatically stored sensor values, most of which are physiological parameters, such as heart rate or oxygen saturation. The rest describes the settings of the medical devices, such as humidity in the incubator or ventilation mode. Despite a large number of different parameters, all of them have not been recorded for all patients as the interest to monitor certain parameters has varied over the years. In addition, different

medical equipment with varying support for parameter monitoring has been used at different times. 14 variables have been recorded for more than 1,000 patients and 34 variables for more than 500 patients. Also, the number of recordings for each parameter varies heavily, and there are 12 parameters with more than 10 million recordings and 32 parameters with more than one million recordings. Table 3.1 presents the parameters that have been recorded automatically for the highest number of patients.

Table 3.1: The most common automatically monitored parameters by the number of patients.

Parameter	Patients	Data entries
Blood oxygen saturation from pulse oximetry	2,053	31,502,272
Heart rate from electrocardiography	2,049	29,388,984
Respiratory rate	2,049	29,271,414
Mean non-invasive blood pressure	1,946	208,760
Systolic non-invasive blood pressure	1,943	199,710
Diastolic non-invasive blood pressure	1,943	199,671
Mean arterial blood pressure	1,923	17,710,660
Systolic arterial blood pressure	1,905	11,967,411
Diastolic arterial blood pressure	1,905	11,967,406
Heart rate from pulse oximetry	1,538	22,037,327
Positive end-expiratory pressure	1,169	9,505,537
Inspiratory:expiratory ratio	1,113	8,777,942
Airway temperature	1,099	14,121,919
Mean airway pressure	1,082	8,643,503
Fraction of inspired oxygen, measured	911	8,549,805
Expiratory tidal volume	911	8,461,850
Lung compliance, measured	904	7,576,377
Fraction of inspired oxygen, set	873	6,715,493
Ventilator respiratory rate	820	7,220,557
Ventilator breath pattern	820	979,357

The medical monitoring devices measure and display sensor data continuously, producing thousands of data recordings for each patient every day (McGregor, 2013). Due to the high price of storage capacity in the late 1990s and early 2000s, the continuous measurements have been stored in a discrete form. At Helsinki University Hospital, they have been discretised so that medians of 10-second-long time intervals have been averaged over two minutes. For a comparison, the discretisation has been calculated as 1-minute averages at Karolinska University Hospital in Stockholm (Honoré, 2017) and as 2-minute medians at Onze Lieve Vrouwe Gasthuis Teaching Hospital in Amsterdam (Bosman et al., 1998).

In addition to automatically stored values, the EHR contains also manually

inserted values for 732 parameters. They are either sensor values, read from the monitor and inserted to the EHR by the hospital personnel, or results of other measurements, such as head circumference or weight of the diapers. These manual measurements have not been recorded with equal time intervals. Similarly to automatically stored values, not all 732 manually monitored parameters have been recorded for all patients but the parameters vary depending on the prevailing practice. In fact, only 152 parameters have been recorded for more than 1,000 patients and 230 for more than 500 patients. There are 18 parameters with more than 200,000 recordings and 36 parameters with more than 100,000 recordings.

The medical diagnoses of the patients have also been stored in the EHR. The diagnosis categorisation follows International statistical classification of diseases and related health problems 10^{th} revision (ICD-10) system (WHO, 2016), and 450 different medical diagnoses have been assigned to the patients. Table 3.2 presents the most common diagnoses.

Other information is contained in the EHR as well. There are 71 different medical procedures, which include, among others, insertion of nasal ventilators for 510 patients, ultrasound imaging of the heart for 57 patients, and nitrogen oxide treatment for 70 patients. Laboratory tests and test results, such as the amount of haemoglobin or leucocytes, are stored in the EHR. Details of the ordered and given medication, such as the volume of saline solution or the amount of medication for diarrhoea, as well as nutrition information, such as the amounts of water, protein and different vitamins, are provided.

Even though the NICU stay of the patients is well documented and many fields of the Clinisoft system are utilised and filled in, there are still many more fields available, which have not been introduced at Helsinki University Hospital. For this study, the relevant part of the data is the basic information of the patients, the automatically monitored parameters, and the diagnoses.

3.1.3 Data quality evaluation

This world-class database is internationally comprehensive as it contains data of 2059 VLBW infants born in 1999–2013. This number corresponds to approximately a third of all VLBW infants born in Finland during those years. The database is proven to be suitable for research as several scientific publications have based their research on that. Rinta-Koski et al. (2017a) and Rinta-Koski et al. (2018) predicted preterm infant mortality and Rinta-Koski et al. (2017b) predicted several morbidities. Immeli et al. (2017) researched the

Table 3.2: The most common diagnoses by the number of patients.

ICD-10 code	Description	Patients
P59.0	Neonatal jaundice associated with preterm delivery	1,107
P07.3	Disorders related to short gestation and low birth weight, not elsewhere classified. Other preterm infants	900
P07.10	Disorders related to short gestation and low birth weight, not elsewhere classified. Other low birth weight (1000–1499 g)	858
P22.9	Respiratory distress of newborn, unspecified	720
P22.0	Respiratory distress syndrome of newborn	672
P07.2	Disorders related to short gestation and low birth weight, not elsewhere classified. Extreme immaturity	527
P29.30	Cardiovascular disorders originating in the perinatal period. Persistent fetal circulation	497
P27.1	Bronchopulmonary dysplasia originating in the perinatal period	416
P00.0	Fetus and newborn affected by maternal hypertensive disorders	388
P05.1	Slow fetal growth and fetal malnutrition. Small for gestational age	373
P01.5	Fetus and newborn affected by multiple pregnancy	359
P07.02	Disorders related to short gestation and low birth weight, not elsewhere classified. Extremely low birth weight (750–999 g)	322
P22.8	Other respiratory distress of newborn	314
P36.3	Sepsis of newborn due to other and unspecified staphylococci	253
P36.90	Bacterial sepsis of newborn, unspecified	227
P07.01	Disorders related to short gestation and low birth weight, not elsewhere classified. Extremely low birth weight (500–749 g)	199
P01.1	Fetus and newborn affected by premature rupture of membranes	198
H35.1	Retinopathy of prematurity	153
P05.0	Slow fetal growth and fetal malnutrition. Light for gestational age	149
P22.1	Transient tachypnoea of newborn	146

postnatal growth of preterm male infants, and Rinta-Koski et al. (2015) the SpO_2 levels of preterm infants.

Despite the comprehensiveness of the database, this EHR contains also challenges that are common to EHRs (see Section 2.5). For example, the automatically recorded measurements are not perfect time series as they contain gaps. The gaps can be caused by misplacement of the sensors, equipment malfunctions, or simply because of the sensors have not been attached to the patient during examinations or washing, for example. In addition, some of the measurement values are clearly out of range. For instance, respiratory rates above 250 breaths per minute or negative values for blood pressure are unacceptable. Moreover, according to the database, 51 patients have entered the NICU before their birth, and physiological variables have been measured from 31 patients before the birth. To point out one more suspiciousness, the

EHR claims there are patients whose gestational ages are more than 4 and 6 years, which is obviously impossible.

Even though the intervals for automatic data recordings should be 2 minutes, the sampling is slightly irregular. There may exist inaccuracies up to a few seconds, which is called irregularity within a time series. It is also remarkable that all parameters for a specific patient are not measured simultaneously but, for example, heart rate may be measured in 2-minute intervals starting from 07:02:15 and SpO_2 in the same intervals starting from 07:02:18. This is called irregularity between the time series.

Finally, the SNAP-II and SNAPPE-II scores have not been defined for all patients due to missing values in the data, resulting in only 1519 and 1023 patients to have them, respectively. Therefore, only a subset of patients can be utilised to build a machine learning model as the scores are included as features in the model.

3.2 Methods

This section encompasses the methodology of this study from data extraction and preprocessing to implementation; this section describes how the data have been transformed to results.

3.2.1 Extracting time series

Four alternative approaches are exploited for preprocessing and extracting the time series from the automatically recorded parameter measurements. Two of them intervene in the issue of irregular sampling within and between the time series, and the other two do not. All four approaches result in distinct time series.

The first preprocessing approach, named *RegAll*, has regular sampling and contains all hours of life from the time interval. It is similar to what has been used by Rinta-Koski et al. (2017a,b, 2018). The extraction of this time series applies the ideas of PAA and LOCF algorithms. First, PAA algorithm is applied to create a time series whose time frames are regular, 2-minute-long and start from the birth. Then, these time frames are assigned with values using LOCF algorithm. The last observed value before the start of each time frame is carried forward to fill the time frames. This way, the algorithm fills the gaps. However, another gap related issue is posed. The same, last observed value is assigned to consecutive time frames during gaps

even though that is not intended. Consequently, this issue is attempted to be tackled by deleting the consecutive, same values, leaving only the first value left, which poses another issue. If the measurement has stayed stable, the algorithm does not recognise that and removes the consecutive values.

The second preprocessing method, *RegExcl6h*, has the same, regular sampling as RegAll but excludes the first six hours of life. The neonatal vital functions are hypothesised to be unstable after birth, which is assumed to produce distorted signals. Therefore, those first hours are omitted.

The third preprocessing technique, named *IrregAll*, contains the irregularly sampled, original time series and all hours of life are included. Similarly to RegExcl6h, the fourth preprocessing approach, *IrregExcl6h*, is a variant of IrregAll where the first six hours of life have been omitted.

3.2.2 Preprocessing the data

Preprocessing the data is essential as it improves the data quality, and thereby, the results of supervised machine learning algorithms (Kotsiantis et al., 2006). All preprocessing steps have been agreed on with medical doctors, neonatologists Prof. S. Andersson (MD, PhD) and M. Leskinen (MD, PhD), at Children's Hospital, University of Helsinki, and Helsinki University Hospital, Helsinki, Finland.

First, all patients, who have died before the 72^{nd} hour of life, are excluded as the process of dying is assumed to affect their physiological signals. The signals are apparently unstable, which would distort the further evaluation. The number of patients excluded at this stage is 59.

Second, this work is interested in identifying the critical patients as early as possible so that medical care can be targeted better to them. It is essential to gain an accurate prediction with a low number of measurements and in a short measuring time (Marlin et al., 2012). Therefore, only the physiological parameter monitoring times starting from the birth and lasting for time periods of 12, 18, 24, 36, 48, and 72 hours are applied. This choice excludes patients who have entered the NICU after the end of those time periods.

Third, the out of range values, discussed earlier in Section 3.1.3, are corrected. On one hand, removing these outliers improves the data quality and minimises their false effect on the results. For example, omitting the values outside of preset limits is a technique to correct the out of range values (Kotsiantis et al., 2006). On the other hand, no information is wanted to be lost by ignoring too many values. Therefore, conservative limits are selected,

which ignore the values that are negative, close-to-zero, and clearly too large, and thus, physiologically impossible. The limits have been set for 14 parameters that have been measured for more than 1,000 patients and they are presented in Table 3.3.

Table 3.3: Lower and upper limits for physiological parameter values.

Physiological parameter	Lower limit	Upper limit
Blood oxygen saturation from pulse oximetry	10	200
Heart rate from electrocardiography	10	250
Respiratory rate	5	250
Mean non-invasive blood pressure	5	200
Systolic non-invasive blood pressure	5	200
Diastolic non-invasive blood pressure	5	200
Mean arterial blood pressure	5	200
Systolic arterial blood pressure	5	200
Diastolic arterial blood pressure	5	200
Heart rate from pulse oximetry	10	350
Positive end-expiratory pressure	0.1	200
Inspiratory:expiratory ratio	0.1	200
Airway temperature	20	50
Mean airway pressure	1	200

Fourth, a subset of the 14 physiological parameters is selected to ensure the comparability of the results to Rinta-Koski et al. (2017a,b, 2018). Five parameters are selected which are blood oxygen saturation, heart rate from electrocardiography as well as mean, systolic, and diastolic arterial blood pressure.

Finally, the problem of incompleteness is tackled. As machine learning algorithms perform well when a sufficient amount of data is available, an adequate number of measurements has to be taken from every patient. That is why a minimum requirement of 50 single measurements for each physiological parameter is employed. Again, this requirement is rather conservative as the automatic measurements are recorded every two minutes, and 50 measurements represent only 3.5 % of their daily maximum number.

As a result, the number of patients after preprocessing is, on average, 925, 951, 958, 964, 968, and 970 for the time periods of 12, 18, 24, 36, 48, and 72 hours, respectively. Table 3.4 presents the exact number of patients and the diagnoses for all four time series and six time intervals. The data are imbalanced for all diagnoses as the proportions of mortality and complications are on a low level.

Table 3.4: The number of patients and diagnoses after preprocessing for all preprocessing and monitoring time combinations.

Pre-processing	Monitoring time	Patients	Mortality	BDP	NEC	ROP
RegAll	12 h	926	60 (6.5 %)	268 (28.9 %)	31 (3.3 %)	73 (7.9 %)
RegAll	18 h	947	63 (6.7 %)	272 (28.7 %)	31 (3.3 %)	75 (7.9 %)
RegAll	24 h	954	63 (6.6 %)	275 (28.8 %)	31 (3.2 %)	77 (8.1 %)
RegAll	36 h	960	63 (6.6 %)	275 (28.6 %)	31 (3.2 %)	77 (8.0 %)
RegAll	48 h	966	63 (6.5 %)	275 (28.5 %)	31 (3.2 %)	77 (8.0 %)
RegAll	72 h	968	63 (6.5 %)	275 (28.4 %)	31 (3.2 %)	77 (8.0 %)
RegExcl6h	12 h	879	58 (6.6 %)	261 (29.7 %)	26 (3.0 %)	70 (8.0 %)
RegExcl6h	18 h	933	62 (6.6 %)	269 (28.8 %)	28 (3.0 %)	74 (7.9 %)
RegExcl6h	24 h	946	62 (6.6 %)	272 (28.8 %)	28 (3.0 %)	76 (8.0 %)
RegExcl6h	36 h	953	62 (6.5 %)	273 (28.6 %)	29 (3.0 %)	76 (8.0 %)
RegExcl6h	48 h	959	62 (6.5 %)	273 (28.5 %)	29 (3.0 %)	76 (7.9 %)
RegExcl6h	72 h	962	62 (6.4 %)	274 (28.5 %)	30 (3.1 %)	77 (8.0 %)
IrregAll	12 h	954	61 (6.4 %)	269 (28.2 %)	31 (3.2 %)	73 (7.7 %)
IrregAll	18 h	967	63 (6.5 %)	273 (28.2 %)	31 (3.2 %)	76 (7.9 %)
IrregAll	24 h	971	63 (6.5 %)	275 (28.3 %)	31 (3.2 %)	77 (7.9 %)
IrregAll	36 h	974	63 (6.5 %)	275 (28.2 %)	31 (3.2 %)	77 (7.9 %)
IrregAll	48 h	977	63 (6.4 %)	275 (28.1 %)	31 (3.2 %)	77 (7.9 %)
IrregAll	72 h	977	63 (6.4 %)	275 (28.1 %)	31 (3.2 %)	77 (7.9 %)
IrregExcl6h	12 h	942	60 (6.4 %)	265 (28.1 %)	28 (3.0 %)	72 (7.6 %)
IrregExcl6h	18 h	956	62 (6.5 %)	269 (28.1 %)	28 (2.9 %)	74 (7.7 %)
IrregExcl6h	24 h	962	62 (6.4 %)	272 (28.3 %)	28 (2.9 %)	76 (7.9 %)
IrregExcl6h	36 h	967	62 (6.4 %)	273 (28.2 %)	29 (3.0 %)	76 (7.9 %)
IrregExcl6h	48 h	971	62 (6.4 %)	274 (28.2 %)	30 (3.1 %)	77 (7.9 %)
IrregExcl6h	72 h	971	62 (6.4 %)	274 (28.2 %)	30 (3.1 %)	77 (7.9 %)

3.2.3 Feature extraction and selection

In feature extraction, two statistical values, mean and standard deviation, are calculated from the preprocessed data of the five physiological time series. These ten features along with the values of GA, BW, SNAP-II, and SNAPPE-II scores are the features of the model, corresponding to Rinta-Koski et al. (2017a,b, 2018). The size of the data matrix is $N \times d$, where N is the number of patients for a specific time series preprocessing and length of monitoring time (see column "Patients" in Table 3.4 for different Ns), and the number of features or dimensions $d = 14$. As the last step, the data are normalised to have a zero mean and unit variance by calculating the z-score for all patients $i = 1, \ldots N$ across all features $j = 1, \ldots, d$ (Duda et al., 2001) as in Equation (3.1),

$$z_j^{(i)} = \frac{x_j^{(i)} - \mu_j}{\sigma_j}, \tag{3.1}$$

where $z_j^{(i)}$ denotes the normalised value, $x_j^{(i)}$ the original value, μ_j and σ_j the feature-specific mean and deviation, respectively.

Furthermore, feature selection is applied manually on the dataset to form four alternate feature combinations as in Rinta-Koski et al. (2018). *TS* means the 10 features derived from the time series, *TS+GA+BW* is the previous along with gestational age and birth weight, *ALL* consists of the previous and SNAP-II and SNAPPE-II scores, and *SC+GA+BW* means the two scores, gestational age, and birth weight.

3.2.4 Implementation

The data preprocessing and classification are implemented in Matlab R2018b (MathWorks, Natick, United States). Additionally, GP models are implemented with `GPstuff` (Vanhatalo et al., 2013), a publicly available toolbox.

All 12 classifiers (introduced in Section 2.3) are applied on the data. They are GP classifiers with four different kernels that are presented in Equations (2.5a), (2.5b), (2.5c), and (2.5d) for linear ($\mathrm{GP}_{\mathrm{linear}}$), Matérn32 ($\mathrm{GP}_{\mathrm{m32}}$), Matérn52 ($\mathrm{GP}_{\mathrm{m52}}$), and RBF ($\mathrm{GP}_{\mathrm{RBF}}$) GP classifiers, respectively. Also NB, LDA, QDA, DT, RF, LR, SVM, and k-NN classifiers are applied. The classification is performed on distinct 96 datasets which represent all combinations of four time series preprocessing alternatives, six monitoring time alternatives, and four feature selection alternatives.

The parameters of two classifiers, k-NN and RF, have been optimised for each complication by grid search. The optimised number of neighbours, k, in k-NN is 16 for mortality, 17 for BPD, 10 for NEC, and 13 for ROP. If k-NN algorithm ends up in a tie between the classes, the label is not selected at random, as explained optionally in Section 2.3.10, but the majority class is selected. The optimised parameters of RF are the number of patients in the leaf nodes and the number of variables that are selected at the splits of the trees. The first parameter values are 8 for mortality, 12 for BPD, and 14 for both NEC and ROP, and the latter values are 2 for mortality and ROP, and 4 for BPD and NEC.

To reduce the statistical uncertainty arising from a relatively small number of patients in the dataset and the split of the data to training and test sets, k-fold cross-validation is applied. It is a common method to, first, split the

dataset into k non-overlapping subsets, and then select $k-1$ of the subsets to train the model, and use the remaining subset to test the model (Goodfellow et al., 2016). This training and testing is repeated k times, so that each of the k subsets are used for testing, one at a time. In fact, stratified k-fold cross-validation is applied, in which the proportion of the classes is equal for each fold. In this work, the 8-fold cross-validation is performed so the sizes of training and test sets are 7/8 and 1/8 of all available data instances, respectively. To reduce the uncertainty of the model even more, the cross-validation has been repeated eight times with different random initialisation. Then, the evaluation measures are calculated as averages over all eight 8-fold cross-validation results, thus being an average of $8 \times 8 = 64$ repetitions.

Finally, the classification performance is assessed by F_1 score and AUPR value due to their suitability for evaluating the imbalanced data as described in Section 2.4. Additionally, AUROC is also reported since it is used commonly in the literature, and it demonstrates how the choice of an inappropriate measure can lead to overly optimistic results.

4. Results

4.1 Optimal classification algorithms

The first research objective is to discover the most appropriate classifiers for mortality and morbidity predictions and to study the differences between the predictability of complications, which is conducted in Section 4.1.1. Additionally, the results are compared to previous studies in Section 4.1.2. To respond to the second goal of this work, the results are reported in less used evaluation measures, F_1 score and AUPR, as well as in a more commonly used measure, AUROC.

4.1.1 Classifier and complication comparison

Reference values

Reference values are set for predictions to see if classifiers outperform these simple prediction techniques. In reference value *Majority*, all patients are assigned with the label of the majority class. That is either not to die or not to be diagnosed. *Random* denotes random guessing of the outcome that is weighted by the class balance. Reference values *SNAP-II* and *SNAPPE-II* use only the respective score to make a prediction by maximising the threshold of the decision boundary subject to accuracy.

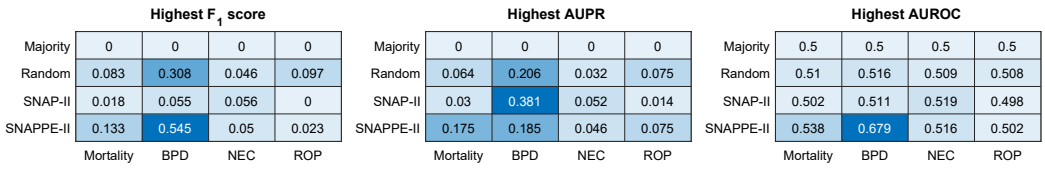


Figure 4.1: Reference performances for mortality, BPD, NEC, and ROP.

Figure 4.1 presents the F_1 scores, AUPRs, and AUROCs for the reference

values. They all remain at a low level for all other outcomes but BPD, indicating their inappropriateness to predict mortality, NEC, or ROP. The medical score SNAPPE-II provides a satisfactory F_1 score of 0.545 for BPD, while SNAP-II results in the highest AUPR of 0.381 for BPD. In terms of AUROC, the threshold of a non-random prediction (0.5) is exceeded clearly only once, implicating that BPD is predictable by SNAPPE-II with an AUROC of 0.679. The majority reference value shows a total unsuitability to predict any of the outcomes.

Predictability of complications of preterm birth

To identify the most suitable classifier for diagnoses predictions, the highest performance values over any of the 96 combinations of time series preprocessing, monitoring time, and feature selection combinations are presented classifier-specifically in Figure 4.2. More comprehensive results with additional performance measures are presented in Appendix A.

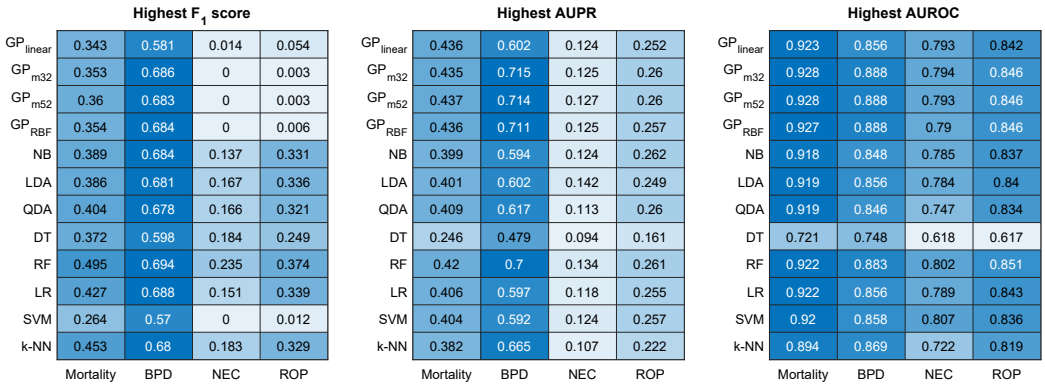


Figure 4.2: Classifier-specific prediction performances for mortality, BPD, NEC, and ROP.

The predictions of BPD reach a high performance in all measures (highest F_1 score: 0.694, AUPR: 0.715, AUROC: 0.888), signifying the potential of being predictable from the data collected at NICUs. Additionally, mortality predictions show a decent performance (highest F_1 score: 0.495, AUPR: 0.437, AUROC: 0.928), whereas the performance of NEC (highest F_1 score: 0.235, AUPR: 0.142, AUROC: 0.807) and ROP (highest F_1 score: 0.374, AUPR: 0.262, AUROC: 0.851) are much weaker in terms of F_1 score and AUPR. Accordingly, the latter two complications are more unpredictable with the procedure employed in this work although their relatively high AUROC values are misleadingly implicating a good predictability. Their low F_1 scores and AUPRs indicate low precision, sensitivity, or both.

Classifier comparison

The preferred classifiers do, indeed, depend on the applied evaluation measure. The random forest classifier performs the best for all outcomes if F_1 score is the criterion. Also k-NN and LR provide a comparable result for most of the outcomes in terms of F_1 score, while NB, LDA, and QDA have more variability in their outcome-specific performance. The GP classifiers show a competitive F_1 score only in BPD predictions.

On the other hand, using AUPR as criterion leads GP to be the most highly performing classifier in mortality (0.715) and BPD (0.437) predictions and one of the best classifiers for NEC (0.127) and ROP (0.260) as well. Furthermore, RF classifier shows a comparable performance for all outcomes. Besides, the remaining classifiers achieve a somewhat lower performance with the exception of DT that produces the poorest AUPRs. The differences in classifier-specific results are rather small for NEC and ROP predictions, and they all remain on a low level.

In case AUROCs are considered, the differences between the classifiers are small. GP, RF, and SVM achieve the highest AUROCs for most of the outcomes, followed closely by the other classifiers except for DT.

4.1.2 Comparison to previous work

Due to the difficulty to compare the results between dissimilar datasets (see Section 2.5 for discussion), a more extensive comparison is conducted to the results of Rinta-Koski et al. (2017b) and Rinta-Koski et al. (2018) as they have investigated the same data with similar research questions. Since precision and sensitivity have been reported in these studies, the corresponding F_1 scores are calculated and used as the evaluation criterion, together with AUROCs. In the comparisons in Table 4.1, the performance is reported for two feature combinations, TS+GA+BW and TS, as only they have been used in all of the previous studies. Since the interest is in comparing the highest achievable prediction performance, the reported F_1 score and AUROC are the highest values over available monitoring lengths. Thus, the lengths may vary between the measures.

None of the classifiers applied in this work is able to outperform the results of the mortality predictions in the previous study by Rinta-Koski et al. (2018). Random forests have the highest F_1 scores: 0.495 for the features of TS+GA+BW and 0.388 for TS. These values are evidently lower than the F_1 scores of Rinta-Koski et al. (2018): 0.524–0.587 for TS+GA+BW and

Table 4.1: Comparison of the results of this study to previous studies of Rinta-Koski et al. (2017b, 2018). Results are reported in two measures: F_1 scores (AUROCs). The highest performance is in bold for each study and both measures.

	Mortality				**Bronchopulmonary dysplasia**			
	TS+GA+BW		TS		TS+GA+BW		TS	
Classifier	R-K[1)]	This study	R-K[1)]	This study	R-K[2)]	This study	R-K[2)]	This study
GP_{linear}	**0.587** (**0.947**)	0.338 (0.906)	0.496 (0.917)	0.212 (0.868)	– (–)	0.581 (0.855)	– (–)	0.502 (0.817)
GP_{m32}	0.554 (**0.947**)	0.338 (0.907)	0.486 (0.925)	0.228 (0.871)	– (–)	0.673 (0.884)	– (–)	0.532 (**0.822**)
GP_{m52}	0.553 (0.946)	0.342 (0.908)	**0.501** (0.925)	0.222 (0.871)	– (–)	0.674 (**0.885**)	– (–)	0.532 (**0.822**)
GP_{RBF}	0.543 (0.946)	0.339 (**0.909**)	0.490 (**0.926**)	0.225 (**0.872**)	**0.59** (**0.87**)	0.677 (0.884)	**0.46** (**0.80**)	0.529 (**0.822**)
NB	– (–)	0.337 (0.897)	– (–)	0.306 (0.864)	– (–)	0.665 (0.847)	– (–)	0.616 (0.812)
LDA	– (–)	0.360 (0.891)	– (–)	0.329 (0.850)	– (–)	0.680 (0.856)	– (–)	**0.630** (0.818)
QDA	– (–)	0.345 (0.831)	– (–)	0.306 (0.808)	– (–)	0.678 (0.846)	– (–)	0.618 (0.799)
DT	– (–)	0.333 (0.674)	– (–)	0.268 (0.630)	– (–)	0.585 (0.727)	– (–)	0.507 (0.676)
RF	– (–)	**0.495** (0.908)	– (–)	**0.388** (0.863)	– (–)	**0.694** (0.877)	– (–)	0.629 (0.817)
LR	– (–)	0.381 (0.893)	– (–)	0.335 (0.868)	– (–)	0.682 (0.856)	– (–)	0.628 (0.817)
SVM	0.524 (0.941)	0.264 (0.894)	0.431 (0.899)	0.157 (0.859)	– (–)	0.570 (0.856)	– (–)	0.493 (0.817)
k-NN	– (–)	0.416 (0.887)	– (–)	0.322 (0.836)	– (–)	0.671 (0.865)	– (–)	0.610 (0.802)

	Necrotising enterocolitis				**Retinopathy of prematurity**			
	TS+GA+BW		TS		TS+GA+BW		TS	
Classifier	R-K[2)]	This study	R-K[2)]	This study	R-K[2)]	This study	R-K[2)]	This study
GP_{linear}	– (–)	0.014 (0.793)	– (–)	0.000 (0.775)	– (–)	0.053 (0.842)	– (–)	0.009 (0.779)
GP_{m32}	– (–)	0.000 (0.794)	– (–)	0.000 (0.764)	– (–)	0.003 (**0.844**)	– (–)	0.003 (0.784)
GP_{m52}	– (–)	0.000 (0.793)	– (–)	0.000 (0.763)	– (–)	0.003 (**0.844**)	– (–)	0.003 (**0.785**)
GP_{RBF}	**0.13** (**0.74**)	0.000 (0.790)	**0.00** (**0.74**)	0.000 (0.757)	**0.09** (**0.84**)	0.006 (**0.844**)	**0.06** (**0.74**)	0.003 (**0.785**)
NB	– (–)	0.121 (0.785)	– (–)	0.106 (0.757)	– (–)	0.327 (0.833)	– (–)	0.265 (0.778)
LDA	– (–)	0.167 (0.784)	– (–)	0.162 (**0.782**)	– (–)	0.336 (0.840)	– (–)	0.268 (0.775)
QDA	– (–)	0.166 (0.747)	– (–)	0.146 (0.740)	– (–)	0.308 (0.782)	– (–)	0.251 (0.729)
DT	– (–)	0.130 (0.567)	– (–)	0.142 (0.578)	– (–)	0.240 (0.602)	– (–)	0.204 (0.576)
RF	– (–)	**0.223** (0.802)	– (–)	**0.232** (0.765)	– (–)	**0.374** (**0.844**)	– (–)	0.298 (0.778)
LR	– (–)	0.151 (0.789)	– (–)	0.143 (0.772)	– (–)	0.339 (0.843)	– (–)	**0.275** (0.779)
SVM	– (–)	0.000 (**0.807**)	– (–)	0.000 (0.765)	– (–)	0.012 (0.836)	– (–)	0.012 (0.778)
k-NN	– (–)	0.171 (0.687)	– (–)	0.183 (0.708)	– (–)	0.329 (0.819)	– (–)	0.262 (0.736)

1) Rinta-Koski et al. (2018), 2) Rinta-Koski et al. (2017b)

0.431–0.501 for TS. The results are lower also in terms of AUROC, but the differences are smaller: the AUROCs of GP and RF classifiers are around 0.04 lower than those of Rinta-Koski et al. (2018). The difference is due to dissimilar data preprocessing. Interestingly, the classification result declines even though the number of patients, and thus the amount of data, was increased from 598 to around 950 in this study. However, this increment simultaneously reduces the mortality rate from 8.8 % to around 6.5 %. As a result, this increase in data imbalance may have a significant impact on the prediction performance.

In BPD predictions, the results are almost the opposite to mortality predictions in terms of F_1 scores. The results of this study are higher in all but three predictions, all of which are modelled with TS+GA+BW features. While the F_1 score of Rinta-Koski et al. (2017b) is 0.59 for TS+GA+BW and 0.46 for TS, many F_1 scores of this work show significantly higher prediction performance: around 0.67–0.69 and 0.61–0.63, respectively. Similarly to mortality predictions, RF, LR, LDA, and QDA belong to the best classifiers. Moreover, the GP classifiers have a comparable performance to the other classifiers if applied on TS+GA+BW features but not if only time series features are used. A reason for the prediction differences between this and the previous studies is the adjustments in the data preprocessing. Another reason may lie in the decreased data imbalance: only 20 % of patients had BPD in the study of Rinta-Koski et al. (2017b), while the rate is around 28 % in this study. Interestingly, all classifiers of this study, except for DT, show almost the same AUROC as the GP_{RBF} classifier of Rinta-Koski et al. (2017b).

In NEC and ROP predictions, both studies share the same data imbalance rate: around 3 % and 7–8 %, respectively, but the absolute number of patients differs due to dissimilar data preprocessing. None of the GP or SVM classifiers shows a satisfactory F_1 score in this study but their AUROCs exceed those of Rinta-Koski et al. (2017b). Apart from RF classifier that reaches F_1 scores of 0.22–0.23 for NEC predictions, the other classifiers have F_1 scores of 0.13–0.18 which are close to the result of the study by Rinta-Koski et al. (2017b), 0.13. Additionally, most of the classifiers of this study receive slightly higher AUROCs than the previous study: 0.75–0.81 in comparison to 0.74 for TS+GA+BW features and 0.76–0.78 in comparison to 0.74 for TS features.

Despite the relatively small difference in NEC predictions between the studies, this study is able to achieve three or four times higher F_1 scores for ROP predictions than the study by Rinta-Koski et al. (2017b). In case TS fea-

tures are combined with GA and BW, the performance is improved from 0.09 to around 0.33 for many classifiers and to 0.37 for random forests. In case without GA and BW, the performance increases from 0.06 to around 0.26 for many classifiers, and RF reaches the highest F_1 score of 0.298. However, the AUROC values of the classifiers do not differ much from the AUROCs of Rinta-Koski et al. (2017b) that are 0.84 for TS+GA+BW features and 0.74 for TS. Only DT stands out with a clearly lower performance of 0.60 and 0.58 for the same feature combinations.

Comparison to other studies

The predictions of this work perform well also in comparison to studies applied on dissimilar data. The mortality predictions are the most successful since the results outperform six out of seven studies. Ramon et al. (2007) received AUROC of 0.88 for NB classification (this study 0.918) and 0.82 for RF (this study 0.922). However, their DT classification with an AUROC of 0.74 outperforms the result of this study, 0.721. Also Salcedo-Bernal et al. (2016) achieved a higher AUROC of 0.74 for DT than this study. However, their results for LR and k-NN, 0.68 and 0.65, respectively, remained lower than the corresponding results of this study that are 0.922 and 0.894, respectively. Finally, SVM classification of this study (AUROC: 0.920) outperformed that of Cerqueira et al. (2014) (AUROC: 0.83).

The complication predictions of this study were not better than existing results. In this study, LR resulted in an AUROC of 0.856 in BPD predictions, while Wajs et al. (2006) achieved an AUROC of 0.94 and Wajs et al. (2007) 0.91. Further, this study predicted ROP and received an F_1 score of 0.374, which is outperformed by the algorithm of Rollins et al. (2015), which achieved an F_1 score of 0.738.

4.2 Optimal data preprocessing and feature selection

This section responds to the third research objective which is to investigate the optimal preprocessing and feature selection approach for predicting neonatal mortality and morbidities. Section 4.2.1 presents the results for the impact of time series preprocessing, Section 4.2.2 for the the impact of the length of monitoring time, and Section 4.2.3 for the impact of feature selection.

4.2.1 Impact of time series preprocessing

Four alternate time series were introduced in Section 3.2.1. Here, the classification results are analysed to reveal if the time series preprocessing affects the performance. The performance is assessed in terms of F_1 score, AUPR, and AUROC, and the highest values over all monitoring time and feature selection combinations are reported in Figures 4.3, 4.4, 4.5, and 4.6 for 12 classifiers. In all figures, the colour bars represent different time series preprocessing, and the value below the name of the classifier is the maximum difference between the performances of the time series preprocessing approaches.

Generally speaking, far-reaching conclusions cannot be drawn from the differences between the distinct time series preprocessing techniques but three obvious observations are to be named. First, the highest difference between the approaches is only 0.034 in F_1 scores, 0.044 in AUPRs, and 0.038 in AUROCs. Moreover, the magnitude of most differences is 0.000–0.015, implicating the different preprocessing has only a minor effect on the performance. Second, the highest performance is not systematically achieved by using a certain preprocessing; one preprocessing functions better for some classifiers and worse for some others. Third, the optimal time series preprocessing depends on the selected evaluation measure; the highest values for all measures are not unambiguously achieved by the same preprocessing approach.

Mortality

Figure 4.3 presents the mortality predictions where the largest number of the highest classification performances is achieved by either RegAll in terms of F_1 score or IrregAll in terms of AUPR and AUROC. This is in favour of not excluding the time series data from the first six hours of life of the infant. However, determining the preferred sampling, regular or irregular, is not fruitful as it depends heavily on the evaluation measure and the classifier. In fact, the choice between regular and irregular sampling is minor: the classification performance changes only by an F_1 score of 0.020, an AUPR of 0.032, and an AUROC of 0.007 if the time series preprocessing is changed from regular to irregular.

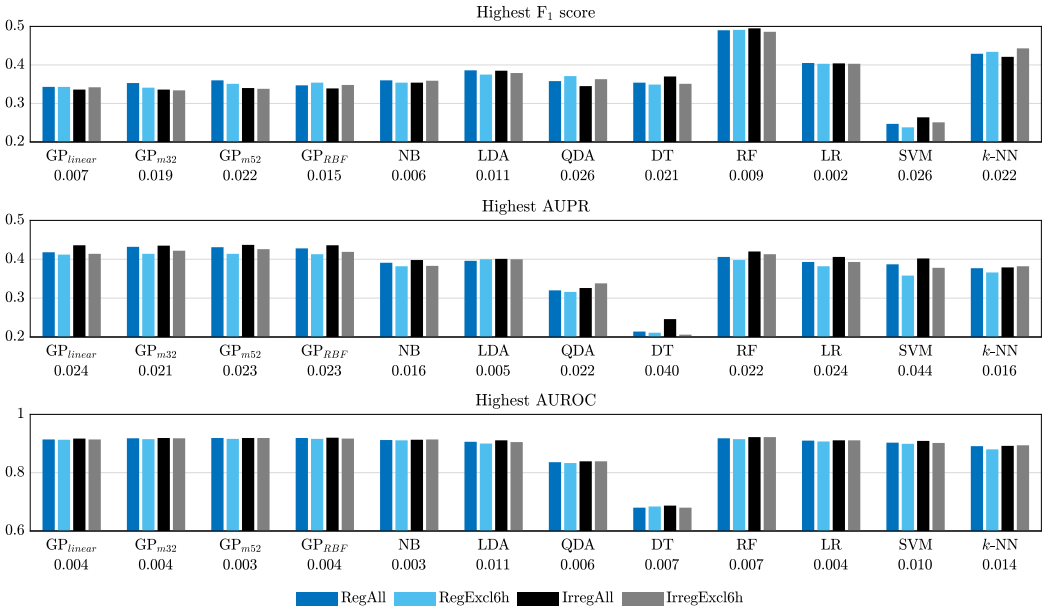


Figure 4.3: Impact of time series preprocessing on the prediction performance of neonatal mortality. The vertical axis is different for AUROC values.
RF, k-NN, and LR dominate the prediction performance in terms of F_1 scores at and above 0.4 whereas GP classifiers achieve the highest AUPRs of around 0.4. AUPRs of GPs are only marginally higher than those of the other classifiers. Only DT shows a low AUPR of around 0.2. All classifiers but QDA and DT result in an almost equal AUROC that is around 0.9.

Bronchopulmonary dysplasia

Time series processing does not affect predicting BPD as much as it affects predicting mortality since the maximum classifier-specific differences (see the values below the names of the classifiers) are smaller for BPD in Figure 4.4 than for mortality in Figure 4.3. The highest difference is 0.015 in F_1 scores, 0.027 in AUPRs, and 0.014 in AUROCs, compared to the respective values of 0.026, 0.044, and 0.014 in mortality predictions. The impact of including or excluding the time series data from the first six hours of life remains controversial due to the negligible differences in the values of the evaluation measures. Nevertheless, the regularly sampled time series, RegAll and RegExcl6h, seem to give, on average, marginally higher performance on both measures than the irregularly sampled time series, IrregAll and IrregExcl6h.

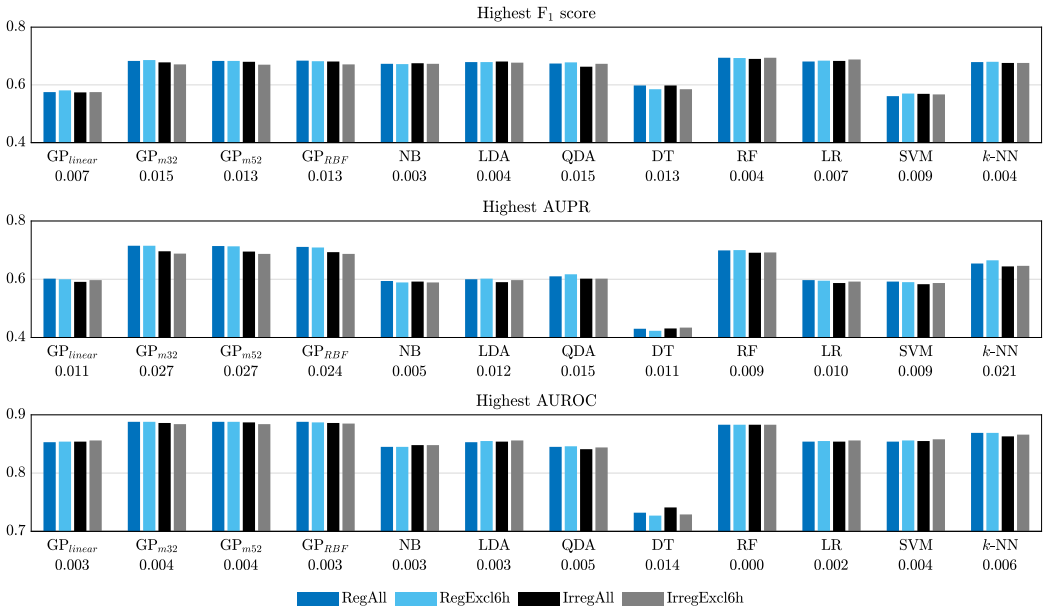


Figure 4.4: Impact of time series preprocessing on the prediction performance of BPD. The vertical axis is different for AUROC values.
Most of the classifiers perform equally well in terms of F_1 score in predicting this lung disease, receiving almost an F_1 score of 0.7. GP and RF classifiers dominate the comparisons of AUPR (at and above 0.7) and AUROC (almost 0.9). The margin to many of the other classifiers is approximately 0.1 in terms of AUPR and even less in terms of AUROC. DT classifier performs significantly worse in terms of AUPR and AUROC.

Necrotising enterocolitis

Figure 4.5 presents the results of NEC predictions. Alike in the mortality and BPD predictions, only minor differences exist between the performance of the different time series preprocessing for NEC. The use of the time series data from the early hours of life does not have any remarkable effect on the performance; the difference is less than 0.01 for most classifiers. Yet, using regularly sampled time series instead of irregularly sampled improves the results a little; there is a positive improvement of AUPR of around 0.010–0.015 for other classifiers than k-NN that receives the highest AUPR with irregular sampling. However, the AUROC values are not affected consistently by the four sampling approaches.

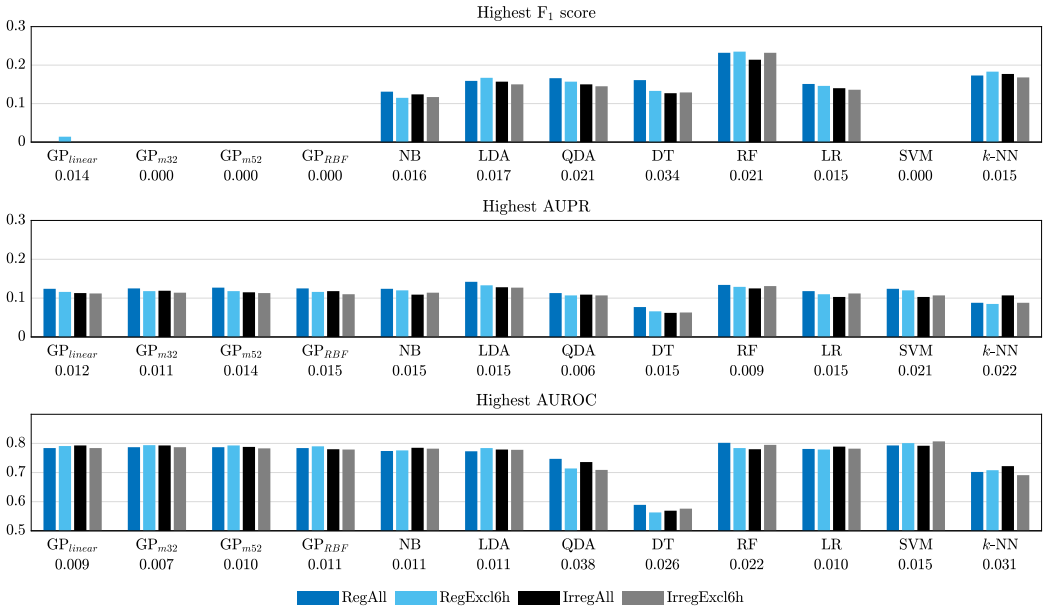


Figure 4.5: Impact of time series preprocessing on the prediction performance of NEC. The vertical axis is different for AUROC values.
Even though the AUROC values are almost 0.8 for other classifiers than DT, QDA, and k-NN, the classifiers are unable to predict NEC: the highest F_1 score is only 0.235 and the following scores are around 0.170. GP and SVM classifiers have a zero result. Additionally, the classification result is equally poor for all classifiers on the AUPR measure; they show an approximate AUPR of 0.1.

Retinopathy of prematurity

The prediction results between the preprocessing approaches barely differ for ROP in Figure 4.6; most of the maximum differences are less than 0.010, and no preprocessing approach shows consistently higher results than any other. Thus, no decent conclusions are drawn about the optimal preprocessing.

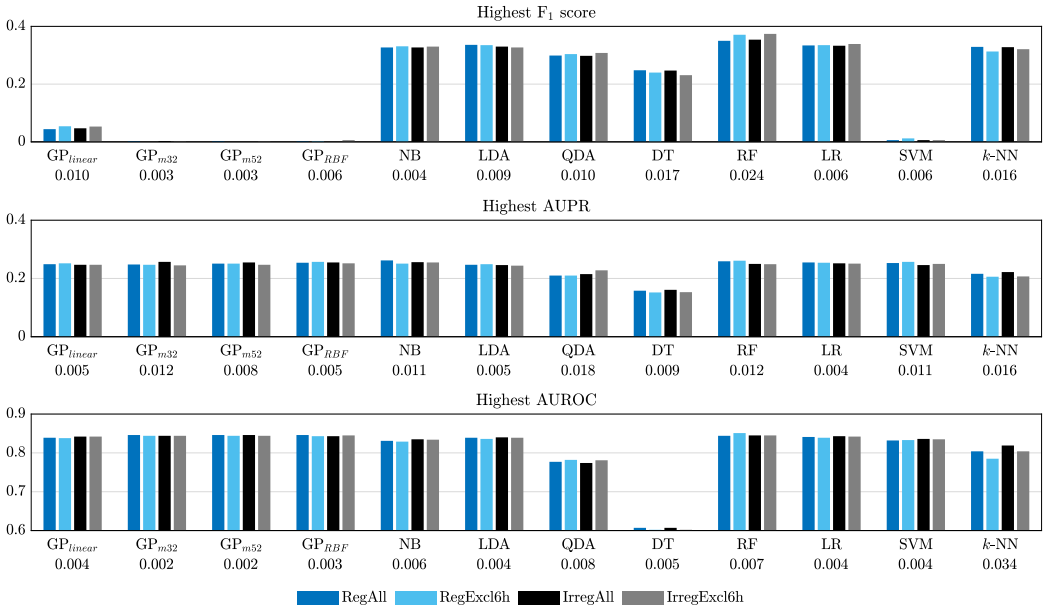


Figure 4.6: Impact of time series preprocessing on the prediction performance of ROP. The vertical axis is different for AUROC values.
RF classifier receives the highest F_1 score also for ROP predictions that is more than 0.35. NB, LDA, LR, and k-NN classifiers show an equal F_1 score of roughly 0.33. Again, the performance is very close to zero for GP and SVM classifiers. In terms of AUPR, all classifiers with the exception of DT are close to an equal performance of 0.25. These low values implicate a poor predictive power for retinopathy even though all other classifiers but QDA and k-NN exceed an AUROC of 0.8.

4.2.2 Impact of the length of the monitoring time

One crucial factor in intensive care is the lack of time. Clinical decisions are to be made as early as possible so that treating the patients can be started before their physical condition deteriorates critically. Therefore, the impact of the length of the physiological parameter monitoring is examined next. In an optimal situation, a satisfactory result is achieved in the shortest conceivable time.

Monitoring times of 12, 18, 24, 36, 48, and 72 hours are of interest. Figures 4.7(a), 4.8, 4.9(a) and 4.10 present the maximum results over all possible time series preprocessing and feature selection combinations, excluding the SC+GA+BW features since they are not time-dependent. The results are barely affected by the length of the monitoring time. Nevertheless, using only TS features shows a difference in mortality and NEC predictions, and that is why they are reported additionally in Figures 4.7(b), and 4.9(b).

Generally speaking, two findings can be made from the comparisons. First, a longer monitoring time includes more information that improves the prediction. The magnitude of this improvement in terms of F_1 score, AUPR, and AUROC remains low for many classifiers – especially in BPD and ROP predictions. Second, the same classifiers tend to have the highest performances regardless of the predicted complication: RF, k-NN, LDA, and LR classifiers are strong if F_1 score is considered, whereas GP and RF classifiers produce many of the highest AUPRs and AUROCs.

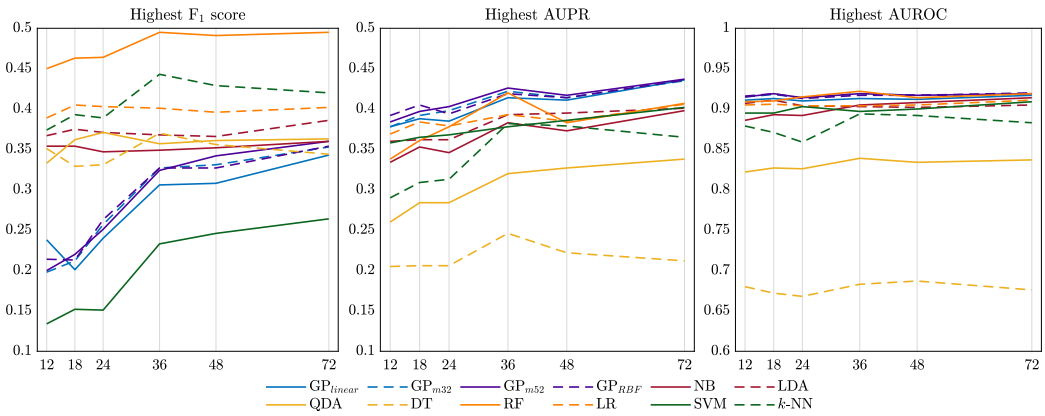
Mortality

Figure 4.7(a) presents the performance of mortality predictions. Measured with F_1 scores, the short monitoring times of 12–24 hours affect the results only a little. Depending on the classifier, the performance increases significantly in the intermediate monitoring times: all GP classifiers improve their performance from approximately 0.2 to 0.3 between the 18^{th} and 36^{th} hours of life, and many other classifiers, such as RF, k-NN, and SVM improve in the interval of 24–36 hours. In the long monitoring times, the performances remain rather stable, and they even begin to decline for a few classifiers, such as k-NN and DT.

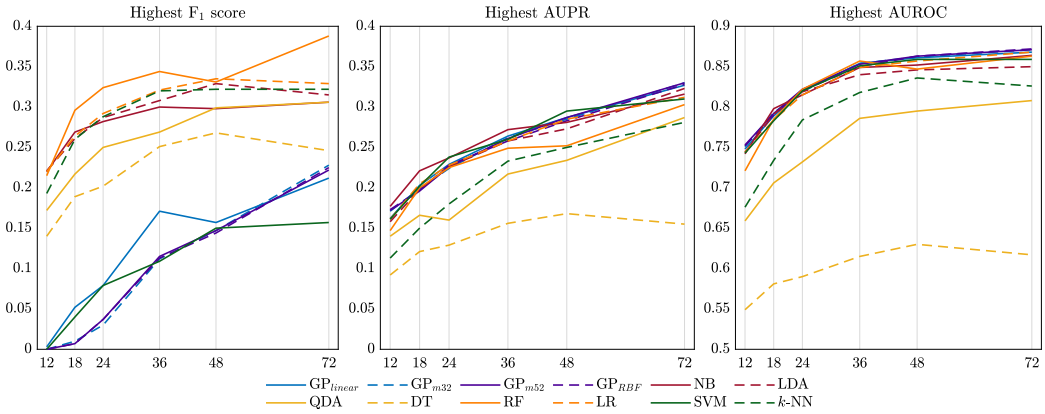
In AUPR evaluations, the performance is also quite steady throughout the monitoring times with the highest growth seen between the 24^{th} and 36^{th} hours of life. Interestingly, most classifiers show a decrease in performance from the 36^{th} to the 48^{th} hour with the exception of LDA and SVM. Their performance keeps on increasing through the entire time period. GP classifiers show the highest AUPRs at around 0.43, followed by RF, LR, SVM, and LDA around 0.035 units behind.

The performance of the classifier is hardly affected by the monitoring time if AUROC is considered. Most of the classifiers are within a narrow margin around the AUROC of 0.9; only QDA and DT perform significantly worse.

Moreover, an interesting and expected phenomenon is visible in Figure 4.7(b) where all classifiers depend solely on the features derived from the physiological time series. The longer monitoring times contribute to the performance



(a) Features include all time-dependent feature combinations.



(b) Features include only the physiological time series.

Figure 4.7: Impact of length of monitoring time on the prediction performance of neonatal mortality. The vertical axis is different for AUROC values.

on both evaluation criteria. The improvement of F_1 scores stabilises for most classifiers at 36 or 48 hours, whereas the GP and SVM classifiers continue to grow over the whole time span. The improvement of AUROC stabilises also after the 36^{th} monitoring hour. Moreover, the AUPRs of all classifiers with an exception of DT, grow significantly from approximately 0.15 to 0.30 in the interval of 12–72 hours. Unfortunately, these results do not answer to the question how long the constant growth would last since monitoring times longer than 72 hours are not studied.

Bronchopulmonary dysplasia

The highest prediction performances of BPD in Figure 4.8 show, basically, no dependency of the length of the monitoring time. The results remain at the same level for all time intervals. All classifiers but SVM, GP_{lin}, and DT have F_1 scores of around 0.65–0.70. On the other hand, GP and RF classifiers stand out at around 0.70 and 0.88 if classifiers are compared by means of AUPR and AUROC, respectively.

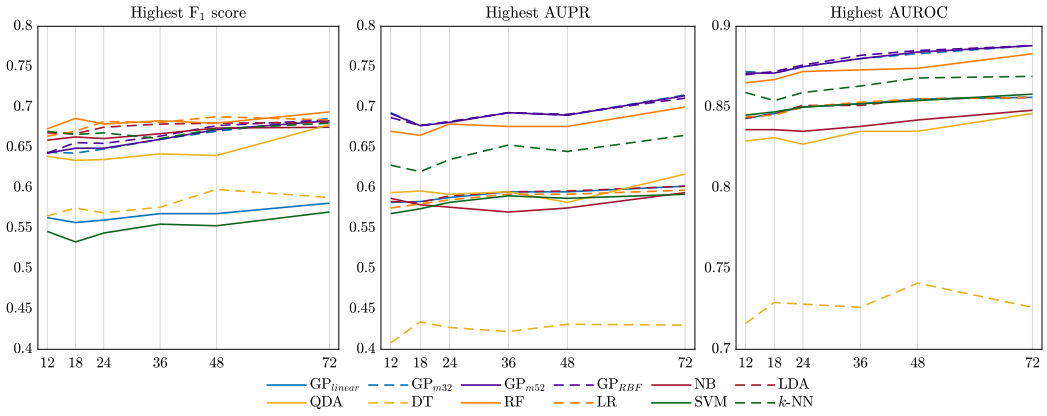
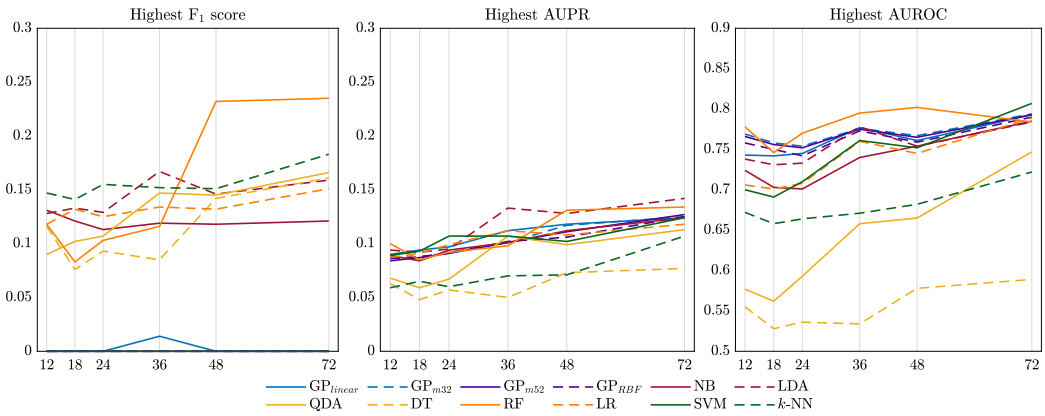


Figure 4.8: Impact of length of monitoring time on the prediction performance of BPD. The vertical axis is different for AUROC values.

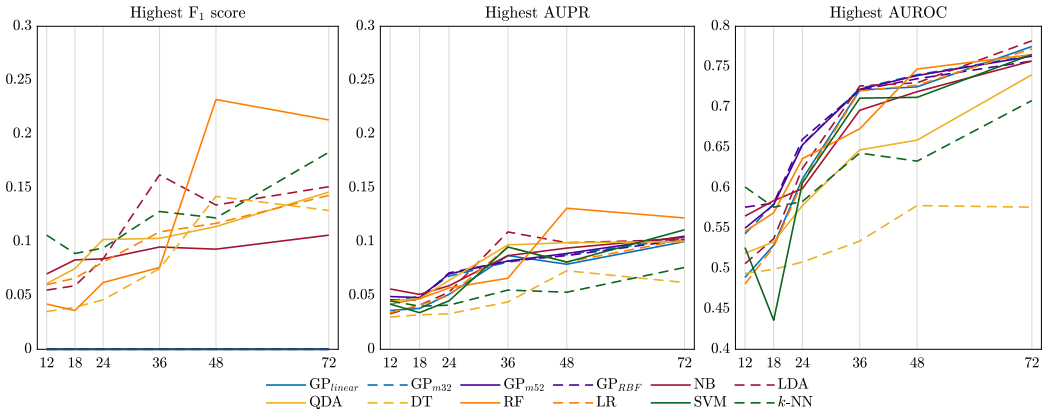
Necrotising enterocolitis

Figure 4.9(a) presents the results for NEC predictions. All in all, the variability is high in the F_1 scores in 12–48 hours, after which most of the classifiers improve their performance. RF and k-NN classifiers achieve the highest F_1 scores, followed by QDA, DT, and LDA. The GP classifiers show a zero in F_1 scores. A clearer pattern is observed in AUPRs as the performance increases almost at a constant rate for the majority of the classifiers over the monitoring times of 12–72 hours. However, the absolute improvement is, on average, only from under 0.100 to roughly 0.125 in terms of AUPR. A small, increasing trend is seen also in the AUROC values, in which QDA shows an ultimate improvement of approximately 0.15.

Using only TS features reveals a similar time-dependency for NEC predictions in Figure 4.9(b) as for mortality predictions in Figure 4.7(b). A longer monitoring time results in higher performance on all measures for all classifiers. A typical magnitude of improvement is 0.05-0.10 in terms of F_1 score and AUPR whereas the growth of AUROC values is roughly 0.20–0.25 for many classifiers. Thus, predicting NEC is highly dependent on the length of



(a) Features include all time-dependent feature combinations.



(b) Features include only the physiological time series.

Figure 4.9: Impact of length of monitoring time on the prediction performance of NEC. The vertical axis is different for AUROC values.

physiological parameter monitoring if they are the only features in the model. In addition, the AUROCs keep on growing throughout the monitored time periods without starting to stabilise.

Retinopathy of prematurity

The attempts to predict retinopathy of prematurity seem to be unaffected by the length of physiological monitoring as the lines are nearly horizontal in all three parts in Figure 4.10. All the same, a proper conclusion is challenging to draw because the level of performance remains low: F_1 scores are at zero for GP and at 0.30–0.35 for other classifiers, and the AUPR values are within a narrow margin around 0.20–0.25 for most of the classifiers. Although AUROC values are much higher, around 0.80–0.85, for ten classifiers, they

implicate an overly optimistic result of the identification of the sick patients because the other values of the measures remain low.

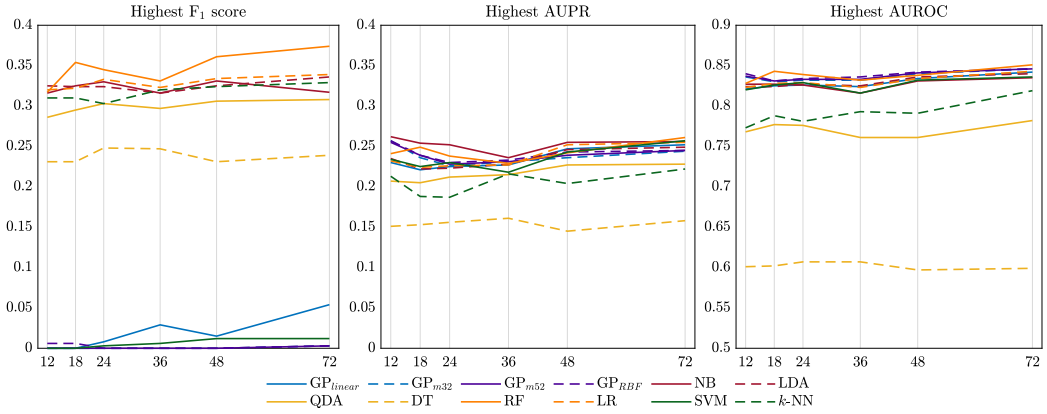


Figure 4.10: Impact of length of monitoring time on the prediction performance of ROP. The vertical axis is different for AUROC values.

4.2.3 Impact of feature selection

The impact of feature selection on the performance is studied in this section. Four proposed feature combinations, ALL, TS+GA+BW, TS, and SC+GA+BW, presented in Section 3.2.3 in more detail, are used to predict neonatal mortality and three morbidities. The highest classification result over all time series preprocessing and monitoring time combinations are presented for each feature combination classifier-specifically in Figures 4.11, 4.12, 4.13, and 4.14. The values below the classifier names are the maximum differences between the performance of the feature combinations.

A few general observations are made from the feature selection comparisons. First, the selected features affect the prediction performance fairly systematically over different classifiers and complications of interest. A high number of features increases the performance. Using only the features derived from time series often results in a clearly lower performance than a combination of them and either GA and BW or the scores SNAP-II and SNAPPE-II, GA, and BW. Second, the four static features in SC+GA+BW alone seem to be very explanatory since they show a strong performance regardless of the evaluation measure. Third, the choice of evaluation measure may alter the optimal feature combination.

Mortality

Figure 4.11 presents the results of mortality predictions. The highest performing feature combinations are ALL, SC+GA+BW, and TS+GA+BW which have only small differences in performance: the highest difference in their F_1 scores is often less than 0.050 and even less on the AUPR and AUROC scales. Thus, using AUPR or AUROC as the target function of the prediction task seems to make the models robust to the selected features. Predicting purely from TS features decreases the performance clearly on all measures.

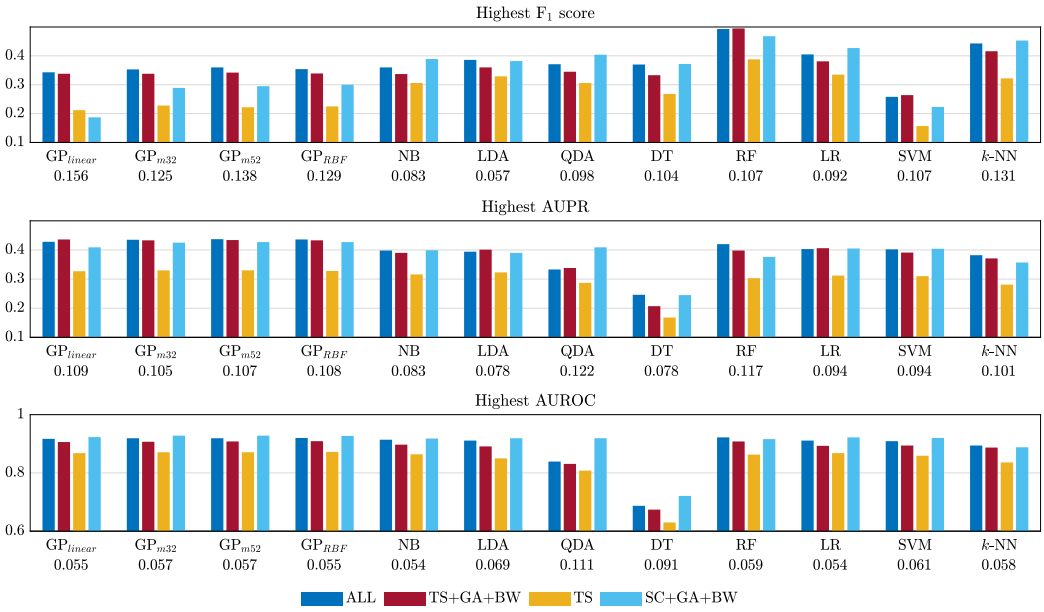


Figure 4.11: Impact of feature selection on the prediction performance of neonatal mortality. The vertical axis is different for AUROC values.
RF, k-NN, and LR classifiers reach the highest results in terms of F_1 score that are 0.495, 0.453, and 0.427, respectively. The other classifiers show F_1 scores of approximately 0.35 while SVM performs significantly worse. Considering AUPR or AUROC, all GP classifiers rank at the top above 0.4 or 0.9, respectively, followed closely by the other classifiers. Only the performance of DT is not comparable to the other classifiers as it only exceeds an AUPR of 0.2 and an AUROC of 0.7.

Bronchopulmonary dysplasia

The results for BPD predictions are presented in Figure 4.12. The impact of feature selection is small among ALL, SC+GA+BW, and TS+GA+BW. In fact, the maximum performance difference within a classifier for these three feature combinations is 0.038 in F_1 scores, 0.067 in AUPRs, and 0.028 in AUROCs. The maximum classifier-specific differences are even smaller between ALL and TS+GA+BW features: 0.013, 0.012, and 0.014 for F_1 score, AUPR, and AUROC, respectively, making these feature combinations almost interchangeable. Predicting this lung disease based on solely TS features decreases the performance remarkably.

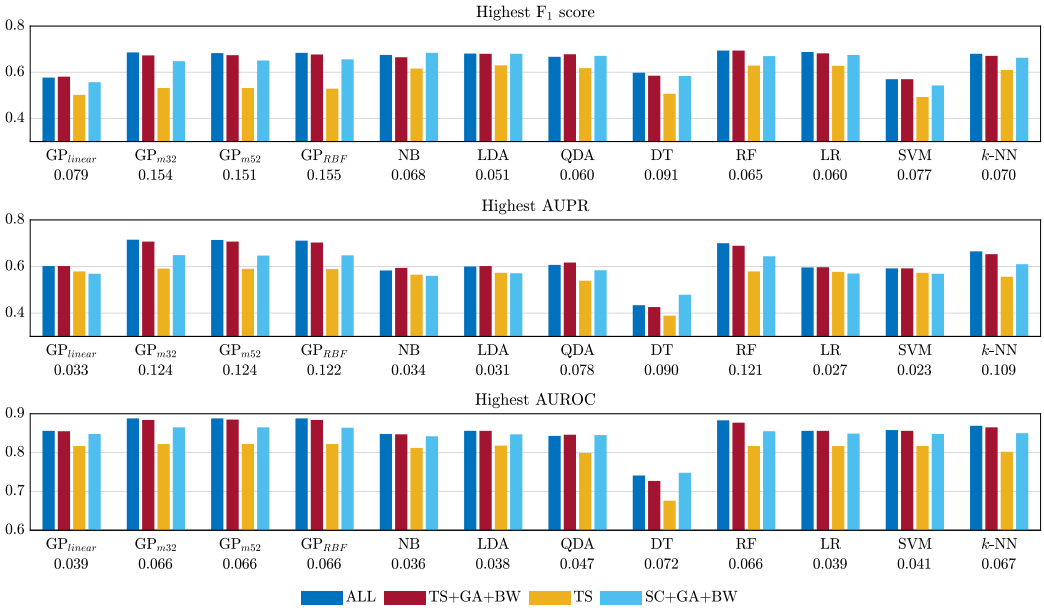


Figure 4.12: Impact of feature selection on the prediction performance of BPD. The vertical axis is different for AUROC values.
All classifiers but DT and SVM achieve almost an equal F_1 score of 0.7. In addition, the GP and RF classifiers hit the same level in terms of AUPR while the majority of the other classifiers show AUPRs of around 0.6. In addition, GP and RF reach almost AUROCs of 0.9, outperforming the other classifiers.

Necrotising enterocolitis

According to the results in Figure 4.13, defining the optimal features is challenging since the differences in classifier-specific performance (see the values below the names of the classifiers) are smaller in NEC predictions than in any other predictions. The maximum differences are 0.076 in F_1 scores, 0.041 in AUPRs, and 0.051 in AUROCs. Unlike in other diagnoses, the TS features perform as well as the other feature combinations. Surprisingly, the performance of TS features only is often comparable to the other feature combinations. Accordingly, the physiological parameters may have a dominant role in predicting NEC, which requires, however, more research due to the low performance of the results at hand.

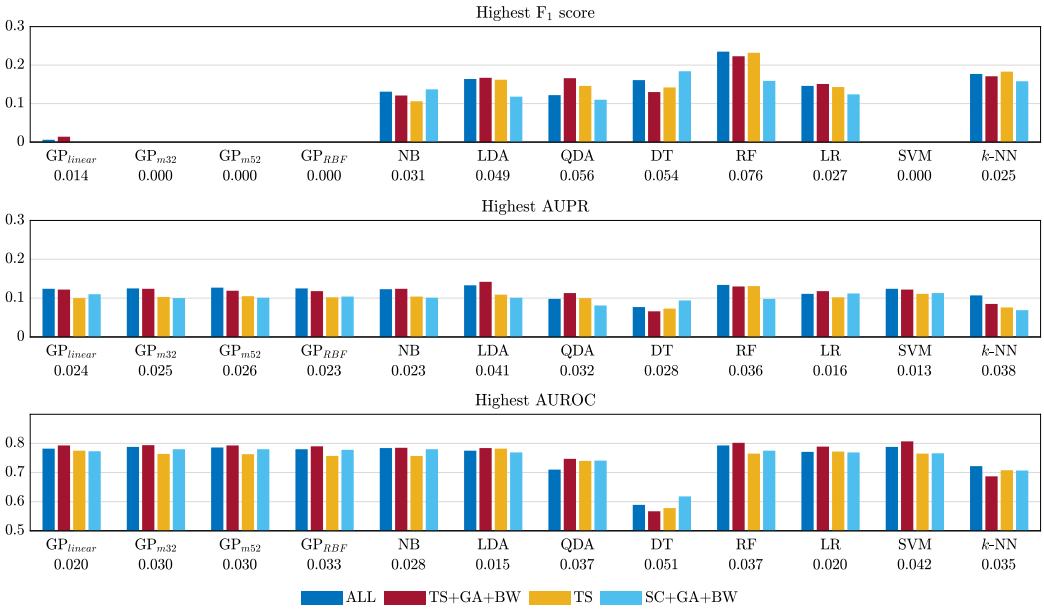


Figure 4.13: Impact of feature selection on the prediction performance of NEC. The vertical axis is different for AUROC values.
Predicting the occurrence of NEC with any feature combination is hard. 11 out of 12 classifiers achieve an approximated AUROC of 0.8, but the F_1 scores barely exceed 0.2 and the values of AUPR remain low at around 0.1 regardless of the classifier and the selected features.

Retinopathy of prematurity

The results for ROP predictions in Figure 4.14 reveal a familiar pattern: TS features alone have lower performance than other feature combinations. The difference is, on average, around 16 %, 10 %, and 6 % lower in terms of F_1 scores, AUPR, and AUROC, respectively. The results of the other feature combinations place within a narrow margin on all measures; the static features of SC+GA+BW usually produce a slightly worse performance than the other two.

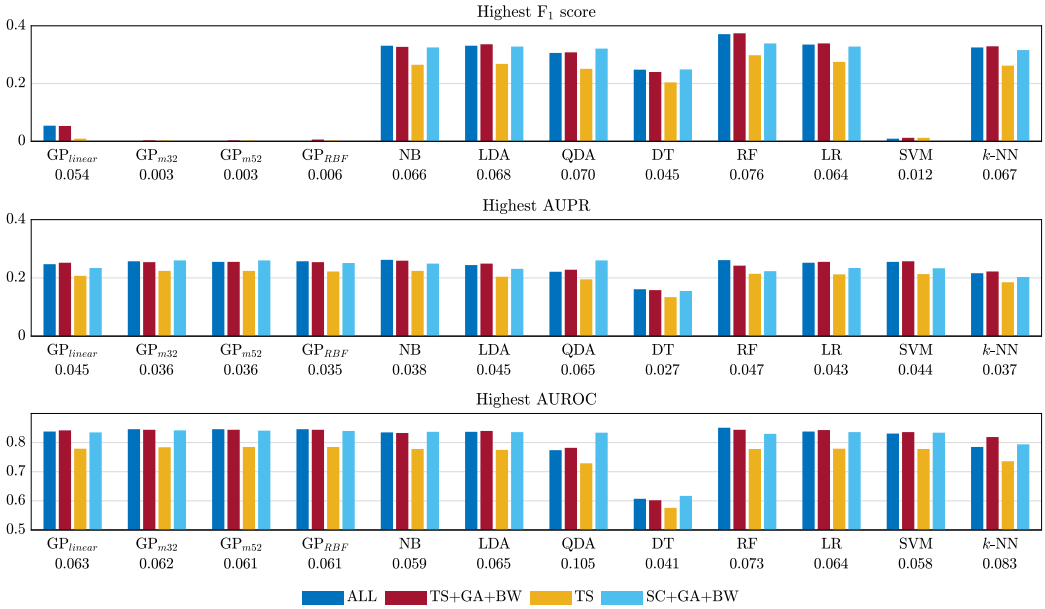


Figure 4.14: Impact of feature selection on the prediction performance of ROP. The vertical axis is different for AUROC values.
The F_1 score is basically zero for GP and SVM classifiers while most of the other classifiers exceed 0.30. The top performance is achieved by RF at around 0.37. The performance is almost equal in terms of AUPR and AUROC: AUPRs are approximately 0.25 and AUROCs exceed 0.80 for all classifiers but DT.

5. Discussion

The first research objective of this work is to identify the most suitable machine learning algorithms to predict neonatal mortality and several complication since no consensus exists in the current literature. The performance of 12 classifiers has been compared, and the preferred classifier depends on the employed performance measure. In terms of F_1 score, the highest result is achieved in most cases with RF classifier, followed by k-NN, and LR classifiers. Only LR has been used widely in the previous research. Also NB, LDA, and QDA show good results in many of the predictions. Optimising the parameters of RF and k-NN may be a reason for their success, especially as the parameters of the other classifiers were not tuned in a similar manner. RF and GP classifiers with either Matérn or RBF kernel result in the highest classification performance in terms of AUPR. RF, GP, and SVM classify the sick patients the best in terms of AUROC. However, any classifier does not clearly outperform the others, verifying the observation of Meyfroidt et al. (2009) that no classifier is more suitable for a certain task than any other.

In addition, the differences in the predictability of complications are of interest in this study. The classification performance presented in Chapter 4 depends heavily on the complication of preterm birth. The highest F_1 scores (AUPRs, AUROCs) are 0.495 (0.437, 0.928) for mortality, 0.694 (0.715, 0.888) for BPD, 0.235 (0.142, 0.807) for NEC, and 0.374 (0.262, 0.851) for ROP. Consequently, BPD of VLBW neonatal infants shows the most potential of being predictable with machine learning since all three measures have a high value. In addition, the results of mortality predictions support the conclusions of previous work by Rinta-Koski et al. (2018) that mortality can be predicted to some extent from physiological parameters.

The results of NEC and ROP predictions remain at a fairly modest level in terms of F_1 score and AUPR, but their AUROCs are close to those of mortality and BPD. That demonstrates clearly the importance of the correct choice of the evaluation criteria. The high AUROC values give an erroneous impression of good predictability of NEC and ROP, but the low F_1 scores

and AUPRs prove the impression to be incorrect. Thus, their potential predictability is not proven with the approaches presented in this work.

The second goal of this study is to present and use appropriate measures for assessing the classification results. At the same time, it explains why NEC and ROP are unpredictable despite their high AUROCs. Even though accuracy, AUC, and AUROC are reported in dozens of studies (see Section 2.6), they are not always the optimal measures. Classifiers function better on balanced than on imbalanced data, and the use of the commonly used measures is justified on balanced data (Weiss and Provost, 2001). However, these evaluation criteria do not measure the relevant aspects on imbalanced data. In the studied patient cohort, the ratio of patients with a diagnosed complication is low; it is around 6.5 % for mortality, 28.5 % for BPD, 3.2 % for NEC, and 7.9 % for ROP.

To present and evaluate the results with more appropriate measures, this thesis applies F_1 score and AUPR. Their origin is in precision and recall. Accordingly, these measures reveal if the classification is able to (i) classify only the truly sick patients as sick and (ii) identify all of the sick patients (Sokolova and Lapalme, 2009; Saito and Rehmsmeier, 2015). However, an acknowledged concern regarding AUPR can slightly mislead the results in this particular study. AUPR applies multiple probability thresholds as explained in Section 2.4.1. The AUPRs of this study are averaged over patient-specific AUPRs, which are not completely equal to the more correct method, gross AUPR. The gross AUPR is calculated using all thresholds for all patients simultaneously without averaging (Ghassemi et al., 2018).

Data imbalance is definitely not the only reason for the low predictability of NEC and ROP because it is higher for mortality (around 6.5 %) than for ROP (around 7.9 %), but still mortality predictions have a significantly higher performance. Another reason for the low performance may lie the feature selection; the selected physiological parameters can cause the low predictability of NEC and ROP if they do not reveal the symptoms of these two complications. For example, many other studies have used retinal images to detect ROP. Of course, a potential source of error can be in the diagnoses that have been given to the patients by the doctors: some patients may potentially been misdiagnosed with a complication, some sick patients may be lacking a diagnosis, or the medical practices have varied over the period of 1999–2013 in a way that certain diagnoses have been given with less evidence at one time than another.

The third research goal is to compare the impact of both data preprocessing, in other words, the time series sampling and the length of monitoring

time, and feature selection on the classification performance. The different preprocessing of the time series has a marginal effect on the results. No generalisations can be drawn whether regularly or irregularly sampled time series are more suitable for the presented analysis, or whether it is advisable to exclude the measurements from the first six hours of life. Moreover, only mean and standard deviation were extracted from the time series, and they are affected only a little by the changes in time series sampling examined in this study. As an important acknowledgement, the choice of preprocessing may have a greater impact on the results if more sophisticated features (see Section 2.2.2) were extracted from the time series.

Furthermore, the effect of the length of the monitoring time is examined as longer time series include more information. The results are twofold: the predictions of mortality and NEC are improved with longer monitoring times, whereas those of BPD and ROP are not affected as remarkably. However, the improvements are moderate: often around 0.05 on all measures. Moreover, somewhat clear improvements in the mortality and NEC predictions are revealed if only features extracted from time series are used in the classification model. An interesting pattern is observed in the varying lengths of the monitoring time: the results are improved the most when the monitoring time is 36 or 48 hours, after which the classification performance usually stabilises. This has also been verified by Rinta-Koski et al. (2018). Therefore, a reasonable monitoring time of the patients is 1.5–2 days. In that time, the most justified predictions of neonatal complications are provided, which can be used to support the decision making at NICUs. A few classifiers show, however, a constant growth in performance throughout the whole 72-hour time period. Since it is the longest monitoring time of this study, it remains unclear how long the growth would last.

The feature selection, on the other hand, affects the results more than data preprocessing. The results of feature selection comparison are similar to those of Saria et al. (2010), Rinta-Koski et al. (2017a), and Rinta-Koski et al. (2018). GA and BW are undoubtedly indispensable features since the highest performance is achieved by the data combinations where they appear. However, it remains unclear if they are to be combined with medical scores, or time series features, or both to achieve the highest results, since the differences between these combinations are minor. One combination happens to function better for a specific classifier and a specific complication, based on which the optimal model can be constructed. All the same, using only time series based features is not advised.

Despite the potential of the predictions proposed in the previous studies and

in this work, most of the medical data analyses end when the numerical results have been analysed, and the results are never implemented in the real life (Bellazzi and Zupan, 2008). This is unfortunate as the algorithms might make a difference at ICUs by improving the quality of care and by saving lives of the newborn. Therefore, the results of a method should not only be tested on one but on multiple patient cohorts to demonstrate their reliability, which is, in most cases, challenging due to the confidential nature of health related data. On top of that, Cerqueira et al. (2014) raise concern about the ethical consequences of medical predictions. If a health care unit has limited resources, is the baby at a higher risk given a priority for the treatment over other patients? The ethical aspects will not be discussed further in this technical thesis.

6. Conclusions

VLBW neonatal infants are prone to multiple medical complications and death due to their underdevelopment and young age. Many of these complications are life-threatening and require immediate care, or at least the treatment is better to be started as early after the birth as possible. Since the physiological condition of preterm infants is monitored continuously with various sensors and manual measurements, neonates produce vast amounts of medical data. Dozens of studies have shown potential of utilising these data by machine learning algorithms. The algorithms can predict the occurrence of typical neonatal complications, thus enabling the doctors to start the proper care in time.

A state-of-the-art NICU patient cohort is used in this study to compare the predictive capability of several classifiers and predictability of different neonatal complications. Random forests, Gaussian processes, k-nearest neighbours, logistic regression, and support vector machine classifiers appear to be the most suitable classifiers for the prediction tasks. The optimal classifier, however, depends on the complication of interest as well as other design choices of the model construction, such as the length of patient monitoring time. This work presents the highest prediction performance for BPD (F_1 score: 0.694, AUPR: 0.715, AUROC: 0.888), followed by decent results for mortality predictions (F_1 score: 0.495, AUPR: 0.437, AUROC: 0.928). NEC (F_1 score: 0.235, AUPR: 0.142, AUROC: 0.807) and ROP (F_1 score: 0.374, AUPR: 0.262, AUROC: 0.851) are not predictable by the proposed technique.

The fortunate rareness of complications is unfortunate from machine learning point of view since the available data are often imbalanced. The class imbalance hinders the use of vanilla machine learning algorithms without a substantial amount of data preprocessing. As an alternative to heavy preprocessing, less used evaluation criteria in this field, F_1 score and AUPR, are utilised since they result in more truthful quantifications of the performance for imbalanced data than many other measures. They focus on assessing the success of identifying the sick patients, not the healthy.

The preprocessing approach of time series appears to be insignificant: regularly and irregularly sampled time series result in almost equal performance. On the other hand, feature selection is more important. Gestational age and birth weight are fundamental features for the model while adding the medical scores SNAP-II and SNAPPE-II or features from the time series might slightly improve the result. Furthermore, this work concludes a longer monitoring time can contribute – depending on the complication – positively to the classification result, suggesting a monitoring time of 36–48 hours. For all that, the effect of feature selection or the length of monitoring time on the classification performance depends on the classifier algorithm and predicted complication.

The concept of successfully predicting neonatal complications using machine learning algorithms receives more evidence in this thesis. However, more research is still required to improve the results of the predictions. Helsinki University Hospital has been collecting neonatal data to a new electronic health record since 2017. This system is able to store continuous patient monitoring values. In consequence, this enables a bunch of new analysis and feature extraction techniques to be applied on the patient cohort. These techniques include heart rate characteristics and beat-to-beat analyses, shapelets, and other more precise temporal features. In addition, the predictive power of other physiological parameters is worth studying as well as more sophisticated feature selection algorithms. Furthermore, the single parameters of the classifiers can be tuned further. All in all, researching EHRs with machine learning algorithms provides work for years to come. This research requires both data science and medical experts so that the quality of care can be improved at NICUs, complications of preterm birth can be healed, and the lives of the neonatal infants can be saved.

Bibliography

Abramowitz, M. and Stegun, I. A. (1965). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover Publications Inc., New York, United States. ISBN 978-0-486-61272-4.

Agrawal, R. and Srikant, R. (1994). Fast Algorithms for Mining Association Rules. In *Proceedings of the 20th International Conference on Very Large Data Bases*, pages 487–499.

Allen, J. F. (1984). Towards a General Theory of Action and Time. *Artificial Intelligence*, 23(2):123–154.

Ambalavanan, N., Van Meurs, K. P., Perritt, R., Carlo, W. A., Ehrenkranz, R. A., Stevenson, D. K., Lemons, J. A., Poole, W. K., and Higgins, R. D. (2008). Predictors of Death or Bronchopulmonary Dysplasia in Preterm Infants with Respiratory Failure. *Journal of Perinatology*, 28(6):420–426.

American Academy of Pediatrics (2004). Age Terminology During the Perinatal Period. *Pediatrics*, 114(5):1362–1364.

Apgar, V. (1953). A Proposal for a New Method of Evaluation of the Newborn Infant. *Current Researches in Anesthesia and Analgesia*, 32(4):260–267.

Ataer-Cansizoglu, E., Bolon-Canedo, V., Campbell, J. P., Bozkurt, A., Erdogmus, D., Kalpathy-Cramer, J., Patel, S., Jonas, K., Chan, R. V. P., Ostmo, S., and Chiang, M. F. (2015). Computer-Based Image Analysis for Plus Disease Diagnosis in Retinopathy of Prematurity: Performance of the "i-ROP" System and Image Features Associated With Expert Diagnosis. *Translational Vision Science & Technology*, 4(6):Article 5.

Avery, G. B., MacDonald, M. G., Seshia, M. M. K., and Mullett, M. D. (2005). *Avery's Neonatology: Pathophysiology & Management of the Newborn*. Lippincott Williams & Wilkins, Philadelphia, United States, 6th edition. ISBN 978-0781746434 (printed), ISBN 978-1469875422 (electronic).

Batal, I., Fradkin, D., Harrison, J., Moerchen, F., and Hauskrecht, M. (2012). Mining Recent Temporal Patterns for Event Detection in Multivariate Time Series Data. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 280–288.

Batal, I., Sacchi, L., Bellazzi, R., and Hauskrecht, M. (2009). Multivariate Time Series Classification with Temporal Abstractions. In *Proceedings of the Twenty-Second International FLAIRS Conference*, pages 344–349.

Bellazzi, R. and Zupan, B. (2008). Predictive data mining in clinical medicine: Current issues and guidelines. *International Journal of Medical Informatics*, 77(2):81–97.

Bhering, C. A., Mochdece, C. C., Moreira, M. E. L., Rocco, J. R., and Sant'Anna, G. M. (2007). Bronchopulmonary dysplasia prediction model for 7-day-old infants. *Jornal de Pediatria*, 83(2):163–170.

Binenbaum, G., Ying, G.-s., Quinn, G. E., Dreiseitl, S., Karp, K., Roberts, R. S., Kirpalani, H., et al. (2011). A Clinical Prediction Model to Stratify Retinopathy of Prematurity Risk Using Postnatal Weight Gain. *Pediatrics*, 127(3):e607–e614.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer, New York, United States. ISBN 978-0387-31073-8.

Bolón-Canedo, V., Ataer-Cansizoglu, E., Erdogmus, D., Kalpathy-Cramer, J., and Chiang, M. F. (2015a). A GMM-based feature extraction technique for the automated diagnosis of Retinopathy of Prematurity. In *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, pages 1498–1501.

Bolón-Canedo, V., Ataer-Cansizoglu, E., Erdogmus, D., Kalpathy-Cramer, J., Fontenla-Romero, O., Alonso-Betanzos, A., and Chiang, M. F. (2015b). Dealing with inter-expert variability in retinopathy of prematurity: A machine learning approach. *Computer Methods and Programs in Biomedicine*, 122(1):1–15.

Bone, R. C., Balk, R. A., Cerra, F. B., Dellinger, R. P., Fein, A. M., Knaus, W. A., Schein, R. M. H., and Sibbald, W. J. (1992). Definitions for Sepsis and Organ Failure and Guidelines for the Use of Innovative Therapies in Sepsis. *Chest*, 101(6):1644–1655.

Bosman, R. J., Oudemans-van Straaten, H. M., and Zandstra, D. F. (1998). The use of intensive care information systems alters outcome prediction. *Intensive Care Medicine*, 24(9):953–958.

Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1):5–32.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification And Regression Trees*. Chapman & Hall/CRC, Boca Raton, United States. ISBN 978-0-412-04841-8.

Calvert, J. S., Price, D. A., Chettipally, U. K., Barton, C. W., Feldman, M. D., Hoffman, J. L., Jay, M., and Das, R. (2016). A computational approach to early sepsis detection. *Computers in Biology and Medicine*, 74:69–73.

Cerqueira, F. R., Ferreira, T. G., de Paiva Oliveira, A., Augusto, D. A., Krempser, E., Barbosa, H. J. C., do Carmo Castro Franceschini, S., de Freitas, B. A. C., Gomes, A. P., and Siqueira-Batista, R. (2014). NICeSim: an open-source simulator based on machine learning techniques to support medical research on prenatal and perinatal care decision making. *Artificial Intelligence in Medicine*, 62(3):193–201.

Chang, C.-C. and Lin, C.-J. (2011). LIBSVM: A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27:1–27:27.

Cirelli, J., McGregor, C., Graydon, B., and James, A. (2013). Analysis of continuous oxygen saturation data for accurate representation of retinal exposure to oxygen in the preterm infant. In Courtney, K. L., Shabestari, O., and Kuo, A., editors, *Enabling Health and Healthcare Through ICT: Available, Tailored and Closer*, pages 126–131. IOS Press, Amsterdam, Netherlands. ISBN 978-1-61499-202-8 (printed). ISBN 978-1-61499-203-5 (electronic).

Cortes, C. and Vapnik, V. (1995). Support-Vector Networks. *Machine Learning*, 20(3):273–297.

Cunha, G. S., Mezzacappa-Filho, F., and Ribeiro, J. D. (2005). Risk Factors for Bronchopulmonary Dysplasia in very Low Birth Weight Newborns Treated with Mechanical Ventilation in the First Week of Life. *Journal of Tropical Pediatrics*, 51(6):334–340.

Darlow, B. A., Hutchinson, J. L., Henderson-Smart, D. J., Donoghue, D. A., Simpson, J. M., and Evans, N. J. (2005). Prenatal Risk Factors for Severe Retinopathy of Prematurity Among Very Preterm Infants of the Australian and New Zealand Neonatal Network. *Pediatrics*, 115(4):990–996.

Davis, J. and Goadrich, M. (2006). The Relationship Between Precision-Recall and ROC

Curves. In *Proceedings of the 23rd International Conference on Machine learning*, pages 233–240.

de Brébisson, A. and Montana, G. (2015). Deep Neural Networks for Anatomical Brain Segmentation. In *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 20–28.

Desautels, T., Calvert, J., Hoffman, J., Jay, M., Kerem, Y., Shieh, L., Shimabukuro, D., Chettipally, U., Feldman, M. D., Barton, C., Wales, D. J., and Das, R. (2016). Prediction of Sepsis in the Intensive Care Unit With Minimal Electronic Health Record Data: A Machine Learning Approach. *JMIR Medical Informatics*, 4(3):e28.

Donders, A. R. T., van der Heijden, G. J. M. G., Stijnen, T., and Moons, K. G. M. (2006). Review: A gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology*, 59(10):1087–1091.

Dorling, J. S., Field, D. J., and Manktelow, B. (2005). Neonatal disease severity scoring systems. *Archives of Disease in Childhood: Fetal and Neonatal Edition*, 90(1):F11–F16.

Duda, R. O., Hart, P. E., and Stork, D. G. (2001). *Pattern Classification*. John Wiley & Sons, Inc., New York, United States, 2nd edition. ISBN 978-0-471-05669-0.

Escobar, G. J. (1999). The Neonatal "Sepsis Work-up": Personal Reflections on the Development of an Evidence-Based Approach Toward Newborn Infections in a Managed Care Organization. *Pediatrics*, 103(Supplement E1):360–373.

Estabrooks, A., Jo, T., and Japkowicz, N. (2004). A Multiple Resampling Method for Learning from Imbalanced Data Sets. *Computational Intelligence*, 20(1):18–36.

Fattore, G., Numerato, D., Peltola, M., Banks, H., Graziani, R., Heijink, R., Over, E., Klitkou, S. T., Fletcher, E., Mihalicza, P., and Sveréus, S. (2015). Variations and Determinants of Mortality and Length of Stay of Very Low Birth Weight and Very Low for Gestational Age Infants in Seven European Countries. *Health Economics*, 24(S2):65–87.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874.

Ferreira, D., Oliveira, A., and Freitas, A. (2012). Applying data mining techniques to improve diagnosis in neonatal jaundice. *BMC Medical Informatics and Decision Making*, 12(1):143.

Ghassemi, M., Pimentel, M. A. F., Naumann, T., Brennan, T., Clifton, D. A., Szolovits, P., and Feng, M. (2015). A Multivariate Timeseries Modeling Approach to Severity of Illness Assessment and Forecasting in ICU with Sparse, Heterogeneous Clinical Data. In *Proceeding of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 446–453.

Ghassemi, M. M., Moody, B. E., Lehman, L. H., Song, C., Li, Q., Sun, H., Mark, R. G., Westover, M. B., and Clifford, G. D. (2018). You Snooze, You Win: the PhysioNet/Computing in Cardiology Challenge 2018. *Computing in Cardiology*, 45:1–4.

Goldberger, A. L., Amaral, L. A., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C.-K., and Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation*, 101(23):E215–E220.

Gomella, T. L., Cunningham, M. D., Eyal, F. G., and Tuttle, D. J. (2013). *Neonatology: Management, Procedures, On-Call Problems, Diseases, and Drugs*. McGraw-Hill Education, New York, United States, 7th edition. ISBN 978-0-07-176801-6 (printed), ISBN 978-0-07-177206-8 (electronic).

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT press, Cambridge, United States. ISBN 978-0-262-03561-3.

Gray, J. E., Richardson, D. K., McCormick, M. C., Workman-Daniels, K., and Goldmann, D. A. (1992). Neonatal therapeutic intervention scoring system: a therapy-based severity-of-illness index. *Pediatrics*, 90(4):561–567.

Griffin, M. P., Lake, D. E., and Moorman, J. R. (2005). Heart Rate Characteristics and Laboratory Tests in Neonatal Sepsis. *Pediatrics*, 115(4):937–941.

Griffin, M. P. and Moorman, J. R. (2001). Toward the Early Diagnosis of Neonatal Sepsis and Sepsis-Like Illness Using Novel Heart Rate Analysis. *Pediatrics*, 107(1):97–104.

Griffin, M. P., O'Shea, T. M., Bissonette, E. A., Harrell, F. E., Lake, D. E., and Moorman, J. R. (2003). Abnormal Heart Rate Characteristics Preceding Neonatal Sepsis and Sepsis-Like Illness. *Pediatric Research*, 53(6):920–926.

Griffin, M. P., O'shea, T. M., Bissonette, E. A., Harrell, F. E., Lake, D. E., and Moorman, J. R. (2004). Abnormal Heart Rate Characteristics Are Associated with Neonatal Mortality. *Pediatric Research*, 55(5):782–788.

Guyon, I. and Elisseeff, A. (2003). An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*, 3(Mar):1157–1182.

Hanley, J. A. and McNeil, B. J. (1982). The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve. *Radiology*, 143(1):29–36.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer Series in Statistics, New York, United States. ISBN 978-0-387-95284-0.

Haukipuro, E.-S., Kolehmainen, V., Myllärinen, J., Remander, S., Salo, J., Takko, T., Nguyen, L. N., Sigg, S., and Findling, R. D. (2019). Mobile Brainwaves: On the Interchangeability of Simple Authentication Tasks with Low-Cost, Single-Electrode EEG Devices. *IEICE Transactions on Communications*, 102(4):760–767.

Häyrinen, K., Saranto, K., and Nykänen, P. (2008). Definition, structure, content, use and impacts of electronic health records: A review of the research literature. *International Journal of Medical Informatics*, 77(5):291–304.

Hellström, A., Smith, L. E. H., and Dammann, O. (2013). Retinopathy of prematurity. *The Lancet*, 382(9902):1445–1457.

Hogan, W. R. and Wagner, M. M. (1997). Accuracy of Data in Computer-based Patient Records. *Journal of the American Medical Informatics Association*, 4(5):342–355.

Holzinger, A. (2016). Interactive machine learning for health informatics: when do we need the human-in-the-loop? *Brain Informatics*, 3(2):119–131.

Honoré, A. (2017). Machine learning for neonatal early warning signs. Master's thesis, KTH Royal Institute of Technology, Stockholm, Sweden. 28 pages.

Immeli, L., Andersson, S., Leskinen, M., Vuorinen, E., Rinta-Koski, O.-P., and Luukkainen, P. (2017). Improved postnatal growth of extremely low-birthweight boys over the last two decades. *Acta Pædiatrica*, 106(4):676–679.

International Neonatal Network (1993). The CRIB (clinical risk index for babies) score: a tool for assessing initial neonatal risk. *The Lancet*, 342(8865):193–198.

ISO/TR 20514:2005(E) (2005). Health informatics – Electronic health record – Definition, scope and context. Geneva, Switzerland, International Organization for Standardization.

Japkowicz, N. and Stephen, S. (2002). The class imbalance problem: A systematic study. *Intelligent Data Analysis*, 6(5):429–449.

Jensen, P. B., Jensen, L. J., and Brunak, S. (2012). Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics*, 13(6):395–405.

Ji, J., Ling, X. B., Zhao, Y., Hu, Z., Zheng, X., Xu, Z., Wen, Q., Kastenberg, Z. J., Li, P., Abdullah, F., Brandt, M. L., Ehrenkranz, R. A., Harris, M. C., Lee, T. C.,

Simpson, J., Bowers, C., Moss, R. L., and Sylvester, K. G. (2014). A data-driven algorithm integrating clinical and laboratory features for the diagnosis and prognosis of necrotizing enterocolitis. *PloS one*, 9(2):e89860.

Keogh, E., Chakrabarti, K., Pazzani, M., and Mehrotra, S. (2001). Dimensionality Reduction for Fast Similarity Search in Large Time Series Databases. *Knowledge and Information Systems*, 3(3):263–286.

Kim, Y. D., Kim, E. A.-R., Kim, K.-S., Pi, S.-Y., and Kang, W. (2005). Scoring Method for Early Prediction of Neonatal Chronic Lung Disease Using Modified Respiratory Parameters. *Journal of Korean Medical Science*, 20(3):397–401.

Knaus, W. A., Draper, E. A., Wagner, D. P., and Zimmerman, J. E. (1985). APACHE II: a severity of disease classification system. *Critical Care Medicine*, 13(10):818–829.

Knaus, W. A., Wagner, D. P., Draper, E. A., Zimmerman, J. E., Bergner, M., Bastos, P. G., Sirio, C. A., Murphy, D. J., Lotring, T., Damiano, A., and Harrell Jr., F. E. (1991). The APACHE III Prognostic System: Risk Prediction of Hospital Mortality for Critically Ill Hospitalized Adults. *Chest*, 100(6):1619–1636.

Knaus, W. A., Zimmerman, J. E., Wagner, D. P., Draper, E. A., and Lawrence, D. E. (1981). APACHE-acute physiology and chronic health evaluation: a physiologically based classification system. *Critical Care Medicine*, 9(8):591–597.

Kotsiantis, S. B., Kanellopoulos, D., and Pintelas, P. E. (2006). Data Preprocessing for Supervised Leaning. *International Journal of Computer Science*, 1(1):111–117.

Kovatchev, B. P., Farhy, L. S., Cao, H., Griffin, M. P., Lake, D. E., and Moorman, J. R. (2003). Sample Asymmetry Analysis of Heart Rate Characteristics with Application to Neonatal Sepsis and Systemic Inflammatory Response Syndrome. *Pediatric Research*, 54(6):892–898.

Laughon, M. M., Langer, J. C., Bose, C. L., Smith, P. B., Ambalavanan, N., Kennedy, K. A., Stoll, B. J., Buchter, S., Laptook, A. R., Ehrenkranz, R. A., Cotten, M. C., Wilson-Costello, D. E., Shankaran, S., Van Meurs, K. P., Davis, A. S., Gantz, M. G., Finer, N. N., Yoder, B. A., Faix, R. G., Carlo, W. A., Schibler, K. R., Newman, N. S., Rich, W., Das, A., Higgins, R. D., and Walsh, M. C. (2011). Prediction of Bronchopulmonary Dysplasia by Postnatal Age in Extremely Premature Infants. *American Journal of Respiratory and Critical Care Medicine*, 183(12):1715–1722.

Le Gall, J.-R., Lemeshow, S., and Saulnier, F. (1993). A New Simplified Acute Physiology Score (SAPS II) Based on a European/North American Multicenter Study. *JAMA*, 270(24):2957–2963.

Le Gall, J. R., Loirat, P., Alperovitch, A., Glaser, P., Granthil, C., Mathieu, D., Mercier, P., Thomas, R., and Villers, D. (1984). A simplified acute physiology score for ICU patients. *Critical Care Medicine*, 12(11):975–977.

Lehman, L. H., Saeed, M., Moody, G. B., and Mark, R. G. (2008). Similarity-Based Searching in Multi-Parameter Time Series Databases. In *Computers in Cardiology*, volume 35, pages 653–656.

Lehman, L.-w. H., Adams, R. P., Mayaud, L., Moody, G. B., Malhotra, A., Mark, R. G., and Nemati, S. (2015). A Physiological Time Series Dynamics-Based Approach to Patient Monitoring and Outcome Prediction. *IEEE Journal of Biomedical and Health Informatics*, 19(3):1068–1076.

Libbrecht, M. W. and Noble, W. S. (2015). Machine learning applications in genetics and genomics. *Nature Reviews Genetics*, 16(6):321–332.

Lin, J., Keogh, E., Wei, L., and Lonardi, S. (2007). Experiencing SAX: a novel symbolic representation of time series. *Data Mining and Knowledge Discovery*, 15(2):107–144.

Löfqvist, C., Andersson, E., Sigurdsson, J., Engström, E., Hård, A.-L., Niklasson, A.,

Smith, L. E. H., and Hellström, A. (2006). Longitudinal Postnatal Weight and Insulin-like Growth Factor I Measurements in the Prediction of Retinopathy of Prematurity. *Archives of Ophthalmology*, 124(12):1711–1718.

Lucas, P. (2004). Bayesian Analysis, Pattern Analysis, and Data Mining in Health Care. *Current Opinion in Critical Care*, 10(5):399–403.

Maier, R. F., Rey, M., Metze, B. C., and Obladen, M. (1997). Comparison of mortality risk: a score for very low birthweight infants. *Archives of Disease in Childhood: Fetal and Neonatal Edition*, 76(3):F146–F151.

Mani, S., Ozdas, A., Aliferis, C., Varol, H. A., Chen, Q., Carnevale, R., Chen, Y., Romano-Keeler, J., Nian, H., and Weitkamp, J.-H. (2014). Medical decision support using machine learning for early detection of late-onset neonatal sepsis. *Journal of the American Medical Informatics Association*, 21(2):326–336.

Marlin, B. M., Kale, D. C., Khemani, R. G., and Wetzel, R. C. (2012). Unsupervised Pattern Discovery in Electronic Health Care Data Using Probabilistic Clustering Models. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, pages 389–398.

Marsland, S. (2015). *Machine Learning: An Algorithmic Perspective*. Chapman & Hall/CRC, Boca Raton, United States, 2nd edition. ISBN 978-1-4665-8333-7.

McGregor, C. (2013). Big Data in Neonatal Intensive Care. *Computer*, 46(6):54–59.

Medlock, S., Ravelli, A. C. J., Tamminga, P., Mol, B. W. M., and Abu-Hanna, A. (2011). Prediction of Mortality in Very Premature Infants: A Systematic Review of Prediction Models. *PloS one*, 6(9):e23441.

Meyfroidt, G., Güiza, F., Ramon, J., and Bruynooghe, M. (2009). Machine learning techniques to examine large patient databases. *Best Practice & Research Clinical Anaesthesiology*, 23(1):127–143.

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill, New York, United States. ISBN 978-0-07-115467-3.

Moorman, J. R., Lake, D. E., and Griffin, M. P. (2006). Heart Rate Characteristics Monitoring for Neonatal Sepsis. *IEEE Transactions on Biomedical Engineering*, 53(1):126–132.

Moskovitch, R. and Shahar, Y. (2015). Classification-driven temporal discretization of multivariate time series. *Data Mining and Knowledge Discovery*, 29(4):871–913.

Murković, I., Steinberg, M. D., and Murković, B. (2003). Sensors in neonatal monitoring: Current practice and future trends. *Technology and Health Care*, 11(6):399–412.

Murphy, K. P. (1998). Switching Kalman Filters. Technical Report 98-10, Compaq Cambridge Research Laboratory, Cambridge, USA.

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. MIT press, Cambridge, United States. ISBN 978-0-262-01802-9.

Nemati, S., Lehman, L.-w. H., Adams, R. P., and Malhotra, A. (2012). Discovering Shared Cardiovascular Dynamics within a Patient Cohort. In *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 6526–6529.

Ochab, M. and Wajs, W. (2014a). Bronchopulmonary Dysplasia Prediction Using Support Vector Machine and LIBSVM. In *Proceedings of the 2014 Federated Conference on Computer Science and Information Systems*, pages 201–208.

Ochab, M. and Wajs, W. (2014b). Bronchopulmonary Dysplasia Prediction Using Support Vector Machine and Logit Regression. In Piętka, E., Kawa, J., and Wieclawek, W., editors, *Information Technologies in Biomedicine, Volume 4. Advances in Intelligent Systems and Computing*, volume 284, pages 365–374. Springer, Cham, Switzerland.

Ochab, M. and Wajs, W. (2016). Expert system supporting an early prediction of the

bronchopulmonary dysplasia. *Computers in Biology and Medicine*, 69:236–244.

Overall, J. E., Tonidandel, S., and Starbuck, R. R. (2009). Last-observation-carried-forward (LOCF) and tests for difference in mean rates of change in controlled repeated measurements designs with dropouts. *Social Science Research*, 38(2):492–503.

Parry, G., Tucker, J., Tarnow-Mordi, W., and UK Neonatal Staffing Study Collaborative Group (2003). CRIB II: an update of the clinical risk index for babies score. *The Lancet*, 361(9371):1789–1791.

Podgorelec, V., Kokol, P., Stiglic, B., and Rozman, I. (2002). Decision Trees: An Overview and Their Use in Medicine. *Journal of Medical Systems*, 26(5):445–463.

Pollack, M. M., Patel, K. M., and Ruttimann, U. E. (1996). PRISM III: an updated Pediatric Risk of Mortality score. *Critical Care Medicine*, 24(5):743–752.

Pollack, M. M., Ruttimann, U. E., and Getson, P. R. (1988). Pediatric risk of mortality (PRISM) score. *Critical Care Medicine*, 16(11):1110–1116.

PostgreSQL Global Development Group (2019). PostgreSQL: The World's Most Advanced Open Source Relational Database. Online. Retrieved from https://www.postgresql.org on 5th March 2019.

Quinn, J. A., Williams, C. K. I., and McIntosh, N. (2009). Factorial Switching Linear Dynamical Systems Applied to Physiological Condition Monitoring. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(9):1537–1551.

Ramon, J., Fierens, D., Güiza, F., Meyfroidt, G., Blockeel, H., Bruynooghe, M., and Van Den Berghe, G. (2007). Mining data from intensive care patients. *Advanced Engineering Informatics*, 21(3):243–256.

Rasmussen, C. E. and Williams, C. K. (2006). *Gaussian Processes for Machine Learning*. The MIT press, Cambridge, United States. ISBN 978-0-262-18253-9.

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

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

Richardson, D. K., Phibbs, C. S., Gray, J. E., McCormick, M. C., Workman-Daniels, K., and Goldmann, D. A. (1993b). Birth Weight and Illness Severity: Independent Predictors of Neonatal Mortality. *Pediatrics*, 91(5):969–975.

Rinta-Koski, O.-P. (2018). Machine learning in neonatal intensive care. Doctoral dissertation. Aalto University, Espoo, Finland. 91+41 pages. ISBN 978-952-60-8209-7 (printed), ISBN 978-952-60-8210-3 (electronic).

Rinta-Koski, O.-P., Hollmén, J., Leskinen, M., and Andersson, S. (2015). Variation in Oxygen Saturation Measurements in Very Low Birth Weight Infants. In *Proceedings of the 8th ACM International Conference on Pervasive Technologies Related to Assistive Environments*, pages 29:1–29:3.

Rinta-Koski, O.-P., Särkkä, S., Hollmén, J., Leskinen, M., and Andersson, S. (2017a). 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*, pages 193–198.

Rinta-Koski, O.-P., Särkkä, S., Hollmén, J., Leskinen, M., and Andersson, S. (2018). Gaussian process classification for prediction of in-hospital mortality among preterm infants. *Neurocomputing*, 298:134–141.

Rinta-Koski, O.-P., Särkkä, S., Hollmén, J., Leskinen, M., Rantakari, K., and Andersson, S. (2017b). Prediction of major complications affecting very low birth weight infants.

In *Proceedings of the 1st IEEE Life Sciences Conference*, pages 186–189.

Rokach, L. (2010). *Pattern Classification Using Ensemble Methods. Series in Machine Perception and Artificial Intelligence*, volume 75. World Scientific, New Jersey, United States. ISBN 978-981-4271-06-6.

Rollins, R., Marshall, A. H., McLoone, E., and Chamney, S. (2015). Discrete conditional phase-type model utilising a multiclass support vector machine for the prediction of retinopathy of prematurity. In *2015 IEEE 28th International Symposium on Computer-Based Medical Systems*, pages 250–255.

Romagnoli, C., Zecca, E., Tortorolo, L., Vento, G., and Tortorolo, G. (1998). A scoring system to predict the evolution of respiratory distress syndrome into chronic lung disease in preterm infants. *Intensive Care Medicine*, 24(5):476–480.

Saar-Tsechansky, M. and Provost, F. (2007). Handling Missing Values when Applying Classification Models. *Journal of Machine Learning Research*, 8(Jul):1623–1657.

Sacchi, L., Larizza, C., Combi, C., and Bellazzi, R. (2007). Data mining with Temporal Abstractions: learning rules from time series. *Data Mining and Knowledge Discovery*, 15(2):217–247.

Saeed, M., Villarroel, M., Reisner, A. T., Clifford, G., Lehman, L.-W., Moody, G., Heldt, T., Kyaw, T. H., Moody, B., and Mark, R. G. (2011). Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II): A public-access intensive care unit database. *Critical Care Medicine*, 39(5):952–960.

Saito, T. and Rehmsmeier, M. (2015). The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PloS one*, 10(3):e0118432.

Salcedo-Bernal, A., Villamil-Giraldo, M. P., and Moreno-Barbosa, A. D. (2016). Clinical data analysis: An opportunity to compare machine learning methods. *Procedia Computer Science*, 100:731–738.

Saria, S., Rajani, A. K., Gould, J., Koller, D., and Penn, A. A. (2010). Integration of Early Physiological Responses Predicts Later Illness Severity in Preterm Infants. *Science Translational Medicine*, 2(48):48ra65.

Seppänen, P., Sund, R., Roos, M., Unkila, R., Meriläinen, M., Helminen, M., Ala-Kokko, T., and Suominen, T. (2016). Obstetric admissions to ICUs in Finland: A multicentre study. *Intensive and Critical Care Nursing*, 35:38–44.

Singer, M., Deutschman, C. S., Seymour, C. W., Shankar-Hari, M., Annane, D., Bauer, M., Bellomo, R., Bernard, G. R., Chiche, J.-D., Coopersmith, C. M., Hotchkiss, R. S., Levy, M. M., Marshall, J. C., Martin, G. S., Opal, S. M., Rubenfeld, G. D., van der Poll, T., Vincent, J.-L., and Angus, D. C. (2016). The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA*, 315(8):801–810.

Sokolova, M. and Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4):427–437.

Stanculescu, I., Williams, C. K. I., and Freer, Y. (2014a). A Hierarchical Switching Linear Dynamical System Applied to the Detection of Sepsis in Neonatal Condition Monitoring. In *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence*, pages 752–761.

Stanculescu, I., Williams, C. K. I., and Freer, Y. (2014b). Autoregressive Hidden Markov Models for the Early Detection of Neonatal Sepsis. *IEEE Journal of Biomedical and Health Informatics*, 18(5):1560–1570.

Stone, M. L., Tatum, P. M., Weitkamp, J.-H., Mukherjee, A. B., Attridge, J., McGahren, E. D., Rodgers, B. M., Lake, D. E., Moorman, J. R., and Fairchild, K. D. (2013). Abnormal heart rate characteristics before clinical diagnosis of necrotizing enterocolitis.

Journal of Perinatology, 33(11):847–850.

Subbe, C. P., Kruger, M., Rutherford, P., and Gemmel, L. (2001). Validation of a modified Early Warning Score in medical admissions. *QJM An International Journal of Medicine*, 94(10):521–526.

Sun, Y., Wong, A. K. C., and Kamel, M. S. (2009). Classification of imbalanced data: A review. *International Journal of Pattern Recognition and Artificial Intelligence*, 23(4):687–719.

Suotsalo, K. and Särkkä, S. (2017). Detecting Malignant Ventricular Arrhythmias in Electrocardiograms by Gaussian Process Classification. In *Proceedings of the 27th IEEE International Workshop on Machine Learning for Signal Processing*, pages 1–5.

Swets, J. A. (1988). Measuring the Accuracy of Diagnostic Systems. *Science*, 240(4857):1285–1293.

Sylvester, K. G., Ling, X. B., Liu, G. Y., Kastenberg, Z. J., Ji, J., Hu, Z., Peng, S., Lau, K., Abdullah, F., Brandt, M. L., Ehrenkranz, R. A., Harris, M. C., Lee, T. C., Simpson, J., Bowers, C., and Moss, R. L. (2014). A novel urine peptide biomarker-based algorithm for the prognosis of necrotising enterocolitis in human infants. *Gut*, 63(8):1284–1292.

Teasdale, G. and Jennett, B. (1974). Assessment of Coma and Impaired Consciousness: A Practical Scale. *The Lancet*, 304(7872):81–84.

Temko, A., Thomas, E., Marnane, W., Lightbody, G., and Boylan, G. (2011). EEG-based neonatal seizure detection with Support Vector Machines. *Clinical Neurophysiology*, 122(3):464–473.

UNICEF, World Health Organization, World Bank Group, and United Nations (2018). Levels & Trends in Child Mortality. Report 2018. Estimates developed by the UN Inter-agency Group for Child Mortality Estimation. Technical report, UNICEF, New York, United States.

United Nations (2019). Neonatal mortality rate (deaths per 1,000 live births). Global SDG Database (online). United Nations, Department of Economic and Social Affairs, Statistics Division. Latest update on the data 21 February 2019. Retrieved from https://unstats.un.org/sdgs/indicators/database/?indicator=3.2.2 on 28th February 2019.

Van Trees, H. L. (1968). *Detection, Estimation, and Modulation Theory – Part I: Detection, Estimation, and Linear Modulation Theory*. John Wiley & Sons, Inc., New York, United States. ISBN 978-0-471-89955-0.

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

Vincent, J.-L., Moreno, R., Takala, J., Willatts, S., De Mendonça, A., Bruining, H., Reinhart, C. K., Suter, P. M., and Thijs, L. G. (1996). The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. *Intensive Care Medicine*, 22(7):707–710.

Wajs, W., Ochab, M., Wais, P., Trojnar, K., and Wojtowicz, H. (2018). Bronchopulmonary Dysplasia Prediction Using Naive Bayes Classifier. In Košcielny, J., Syfert, M., and Sztyber, A., editors, *Advanced Solutions in Diagnostics and Fault Tolerant Control. DPS 2017. Advances in Intelligent Systems and Computing*, volume 635, pages 281–290. Springer, Cham, Switzerland.

Wajs, W., Stoch, P., and Kruczek, P. (2006). Bronchopulmonary Dysplasia Prediction using Logistic Regression. In *Sixth International Conference on Intelligent Systems Design and Applications*, volume 3, pages 98–102.

Wajs, W., Stoch, P., and Kruczek, P. (2007). Radial Basis Networks and Logistic Regres-

sion Method for Prediction of Bronchopulmonary Dysplasia. In *Seventh International Conference on Intelligent Systems Design and Applications*, pages 551–555.

Walsh, M. C., Szefler, S., Davis, J., Allen, M., Van Marter, L., Abman, S., Blackmon, L., and Jobe, A. (2006). Summary Proceedings from the Bronchopulmonary Dysplasia Group. *Pediatrics*, 117(Supplement 1):S52–S56.

Wang, K., Bhandari, V., Chepustanova, S., Huber, G., O'Hara, S., O'Hern, C. S., Shattuck, M. D., and Kirby, M. (2013). Which Biomarkers Reveal Neonatal Sepsis? *PloS one*, 8(12):e82700.

Weiss, G. M. and Provost, F. (2001). The Effect of Class Distribution on Classifier Learning: An Empirical Study. Technical Report ML-TR-44, Department of Computer Science, Rutgers University, New Brunswick, USA.

WHO (2016). *International statistical classification of diseases and related health problems – 10th revision*. World Health Organization, Geneva, Switzerland, 5th edition.

WHO (2018). Preterm birth. Fact sheet. World Health Oraganization, Geneva, Switzerland. Retrieved from https://www.who.int/news-room/fact-sheets/detail/preterm-birth on 31st January 2019.

WHO and MCEE (2018). MCEE-WHO methods and data sources for child causes of death 2000–2017. Number of deaths among children under 5 years by cause, 2017. Technical report, World Health Organization & Maternal and Child Epidemiology Estimation Group, Geneva, Switzerland.

Wittenberg, L. A., Jonsson, N. J., Chan, R. V. P., and Chiang, M. F. (2012). Computer-Based Image Analysis for Plus Disease Diagnosis in Retinopathy of Prematurity. *Journal of Pediatric Ophthalmology and Strabismus*, 49(1):11–19.

Wold, H. (1938). *A Study in the Analysis of Stationary Time Series*. PhD thesis, Almqvist & Wiksell, Uppsala, Sweden.

Wu, C., Löfqvist, C., Smith, L. E. H., VanderVeen, D. K., and Hellström, A. (2012). Importance of Early Postnatal Weight Gain for Normal Retinal Angiogenesis in Very Preterm Infants: A Multicenter Study Analyzing Weight Velocity Deviations for the Prediction of Retinopathy of Prematurity. *Archives of Ophthalmology*, 130(8):992–999.

Ye, L. and Keogh, E. (2009). Time Series Shapelets: A New Primitive for Data Mining. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 947–956.

Zhao, J., Papapetrou, P., Asker, L., and Boström, H. (2017). Learning from heterogeneous temporal data in electronic health records. *Journal of Biomedical Informatics*, 65:105–119.

Zimmerman, J. E., Kramer, A. A., McNair, D. S., and Malila, F. M. (2006). Acute Physiology and Chronic Health Evaluation (APACHE) IV: Hospital mortality assessment for today's critically ill patients. *Critical Care Medicine*, 34(5):1297–1310.

A. Highest classification results

The classification results over all possible time series preprocessing, monitoring time, and feature selection combinations are presented complication- and classifier-specifically in Appendix A. For each complication and classifier combination, the highest F_1 scores are reported in Table A.1, the highest AUPR values in Table A.2, and the highest AUROC values in Table A.3.

If several combinations share the exactly same performance in terms of F_1 score in Table A.1, AUPR in Table A.2, or AUROC in Table A.3, the combination with the highest performance of the secondary measure, AUPR, F_1 score, and F_1 score, respectively, is presented in the table and highlighted with an asterisk (*).

In case several combinations have the exactly same performance in terms of the two measures, the combination with the highest performance of the tertiary measure, AUROC, AUROC, and AUPR, respectively, is presented in the table and highlighted with a dagger (†).

If multiple combinations have the same performance in all three scores, they all are presented in the table and highlighted with a double dagger (‡).

Table A.1: Highest classification results complication- and classifier-specifically in terms of F_1 scores

Complication	Classifier	Pre-processing	Feature selection	Monitoring time (h)	Accuracy (SE)	Precision (SE)	Sensitivity (SE)	Specificity (SE)	F_1 score (SE)	AUPR (SE)	AUROC (SE)
Mortality	GP_{linear} (*)	RegAll	ALL	72	0.940 (0.00)	0.644 (0.03)	0.253 (0.02)	0.988 (0.00)	0.343 (0.02)	0.415 (0.02)	0.914 (0.00)
Mortality	GP_{m32}	RegAll	ALL	72	0.942 (0.00)	0.683 (0.03)	0.259 (0.02)	0.989 (0.00)	0.353 (0.02)	0.430 (0.02)	0.918 (0.01)
Mortality	GP_{m52}	RegAll	ALL	72	0.942 (0.00)	0.692 (0.03)	0.263 (0.02)	0.990 (0.00)	0.360 (0.02)	0.430 (0.02)	0.919 (0.01)
Mortality	GP_{RBF}	RegExcl6h	ALL	72	0.943 (0.00)	0.666 (0.04)	0.259 (0.02)	0.990 (0.00)	0.354 (0.02)	0.410 (0.02)	0.914 (0.01)
Mortality	NB	RegAll	SC+GA+BW	18	0.808 (0.00)	0.249 (0.01)	0.907 (0.01)	0.801 (0.00)	0.389 (0.01)	0.397 (0.02)	0.915 (0.01)
Mortality	LDA	RegAll	ALL	72	0.828 (0.00)	0.254 (0.01)	0.820 (0.02)	0.828 (0.00)	0.386 (0.01)	0.388 (0.02)	0.903 (0.01)
Mortality	QDA	IrregAll	SC+GA+BW	18	0.822 (0.00)	0.262 (0.01)	0.910 (0.01)	0.816 (0.00)	0.404 (0.01)	0.390 (0.01)	0.919 (0.00)
Mortality	DT	IrregAll	SC+GA+BW	24	0.889 (0.00)	0.303 (0.01)	0.502 (0.02)	0.916 (0.00)	0.372 (0.01)	0.234 (0.01)	0.714 (0.01)
Mortality	RF (*)	IrregAll	TS+GA+BW	72	0.909 (0.00)	0.399 (0.01)	0.682 (0.02)	0.924 (0.00)	**0.495** (0.01)	0.389 (0.02)	0.907 (0.01)
Mortality	LR (*)	IrregAll	SC+GA+BW	18	0.839 (0.00)	0.283 (0.01)	0.900 (0.01)	0.835 (0.01)	0.427 (0.01)	0.397 (0.01)	0.922 (0.00)
Mortality	SVM	IrregAll	TS+GA+BW	72	0.944 (0.00)	0.788 (0.04)	0.170 (0.02)	0.997 (0.00)	0.264 (0.02)	0.391 (0.02)	0.894 (0.01)
Mortality	*k*-NN	IrregAll	SC+GA+BW	18	0.869 (0.00)	0.317 (0.01)	0.812 (0.02)	0.873 (0.00)	0.453 (0.01)	0.341 (0.01)	0.885 (0.01)
BPD	GP_{linear}	RegExcl6h	TS+GA+BW	72	0.785 (0.00)	0.653 (0.01)	0.527 (0.01)	0.887 (0.00)	0.581 (0.01)	0.600 (0.01)	0.854 (0.00)
BPD	GP_{m32}	RegExcl6h	ALL	72	0.823 (0.00)	0.696 (0.01)	0.680 (0.01)	0.879 (0.00)	0.686 (0.01)	0.715 (0.01)	0.888 (0.00)
BPD	GP_{m52} (*)	RegAll	ALL	72	0.822 (0.00)	0.697 (0.01)	0.675 (0.01)	0.881 (0.00)	0.683 (0.01)	0.714 (0.01)	0.888 (0.00)
BPD	GP_{RBF}	RegAll	ALL	72	0.822 (0.00)	0.693 (0.01)	0.680 (0.01)	0.878 (0.00)	0.684 (0.01)	0.711 (0.01)	0.888 (0.00)
BPD	NB	IrregAll	SC+GA+BW	24	0.791 (0.00)	0.603 (0.01)	0.794 (0.01)	0.791 (0.01)	0.684 (0.01)	0.555 (0.01)	0.841 (0.00)
BPD	LDA	IrregAll	ALL	72	0.789 (0.00)	0.595 (0.01)	0.798 (0.01)	0.786 (0.00)	0.681 (0.01)	0.589 (0.01)	0.854 (0.00)
BPD	QDA	RegExcl6h	TS+GA+BW	72	0.792 (0.00)	0.608 (0.01)	0.769 (0.01)	0.801 (0.01)	0.678 (0.01)	0.617 (0.01)	0.846 (0.00)
BPD	DT (*)	IrregAll	ALL	48	0.758 (0.00)	0.564 (0.01)	0.643 (0.01)	0.802 (0.01)	0.598 (0.01)	0.431 (0.01)	0.741 (0.01)
BPD	RF (*)	RegAll	ALL	72	0.796 (0.00)	0.608 (0.01)	0.811 (0.01)	0.790 (0.01)	**0.694** (0.01)	0.699 (0.01)	0.883 (0.00)
BPD	LR	IrregExcl6h	ALL	48	0.794 (0.00)	0.602 (0.01)	0.806 (0.01)	0.789 (0.00)	0.688 (0.01)	0.585 (0.01)	0.855 (0.00)
BPD	SVM (*)	RegExcl6h	TS+GA+BW	72	0.783 (0.00)	0.654 (0.01)	0.509 (0.01)	0.892 (0.00)	0.570 (0.01)	0.587 (0.01)	0.853 (0.00)
BPD	*k*-NN	RegExcl6h	ALL	72	0.769 (0.00)	0.564 (0.01)	0.860 (0.01)	0.733 (0.01)	0.680 (0.01)	0.665 (0.01)	0.869 (0.00)
NEC	GP_{linear}	RegExcl6h	TS+GA+BW	36	0.968 (0.00)	0.828 (0.05)	0.009 (0.01)	0.998 (0.00)	0.014 (0.01)	0.110 (0.01)	0.762 (0.01)
NEC	GP_{m32} (*)	RegAll	ALL	72	0.968 (0.00)	1.000 (0.00)	0.000 (0.00)	1.000 (0.00)	0.000 (0.00)	0.125 (0.01)	0.781 (0.01)
NEC	GP_{m52} (*)	RegAll	ALL	72	0.968 (0.00)	1.000 (0.00)	0.000 (0.00)	1.000 (0.00)	0.000 (0.00)	0.127 (0.01)	0.780 (0.01)
NEC	GP_{RBF} (*)	RegAll	ALL	72	0.968 (0.00)	1.000 (0.00)	0.000 (0.00)	1.000 (0.00)	0.000 (0.00)	0.125 (0.01)	0.778 (0.01)
NEC	NB (‡)	IrregAll	SC+GA+BW	48	0.695 (0.01)	0.075 (0.00)	0.749 (0.03)	0.693 (0.01)	0.137 (0.01)	0.096 (0.01)	0.779 (0.01)
NEC	NB (‡)	IrregAll	SC+GA+BW	72	0.695 (0.01)	0.075 (0.00)	0.749 (0.03)	0.693 (0.01)	0.137 (0.01)	0.096 (0.01)	0.779 (0.01)
NEC	LDA	RegExcl6h	TS+GA+BW	36	0.813 (0.00)	0.097 (0.01)	0.611 (0.03)	0.820 (0.00)	0.167 (0.01)	0.132 (0.01)	0.756 (0.02)
NEC	QDA	RegAll	TS+GA+BW	72	0.859 (0.00)	0.103 (0.01)	0.449 (0.03)	0.873 (0.00)	0.166 (0.01)	0.113 (0.01)	0.747 (0.02)
NEC	DT (‡)	IrregAll	SC+GA+BW	48	0.920 (0.00)	0.139 (0.01)	0.296 (0.03)	0.940 (0.00)	0.184 (0.02)	0.094 (0.01)	0.618 (0.02)
NEC	DT (‡)	IrregAll	SC+GA+BW	72	0.920 (0.00)	0.139 (0.01)	0.296 (0.03)	0.940 (0.00)	0.184 (0.02)	0.094 (0.01)	0.618 (0.02)
NEC	RF	RegExcl6h	ALL	72	0.943 (0.00)	0.224 (0.02)	0.289 (0.03)	0.964 (0.00)	**0.235** (0.02)	0.129 (0.01)	0.784 (0.01)
NEC	LR	RegAll	TS+GA+BW	72	0.748 (0.00)	0.085 (0.00)	0.694 (0.03)	0.749 (0.01)	0.151 (0.01)	0.118 (0.01)	0.781 (0.01)
NEC	SVM (*)	RegAll	ALL	72	0.968 (0.00)	1.000 (0.00)	0.000 (0.00)	1.000 (0.00)	0.000 (0.00)	0.124 (0.01)	0.782 (0.01)
NEC	*k*-NN	RegExcl6h	TS	72	0.840 (0.00)	0.110 (0.01)	0.573 (0.03)	0.849 (0.00)	0.183 (0.01)	0.073 (0.00)	0.708 (0.01)
ROP	GP_{linear}	RegExcl6h	ALL	72	0.919 (0.00)	0.667 (0.06)	0.031 (0.01)	0.996 (0.00)	0.054 (0.01)	0.247 (0.01)	0.834 (0.01)
ROP	GP_{m32} (*)	IrregAll	TS+GA+BW	72	0.921 (0.00)	1.000 (0.00)	0.002 (0.00)	1.000 (0.00)	0.003 (0.00)	0.236 (0.01)	0.843 (0.01)
ROP	GP_{m52} (*)	RegAll	TS+GA+BW	72	0.921 (0.00)	1.000 (0.00)	0.002 (0.00)	1.000 (0.00)	0.003 (0.00)	0.245 (0.01)	0.844 (0.01)
ROP	GP_{RBF} (*)	IrregExcl6h	TS+GA+BW	12	0.924 (0.00)	0.977 (0.02)	0.003 (0.00)	1.000 (0.00)	0.006 (0.00)	0.233 (0.01)	0.832 (0.01)
ROP	NB	RegExcl6h	ALL	48	0.742 (0.01)	0.210 (0.01)	0.794 (0.02)	0.738 (0.01)	0.331 (0.01)	0.237 (0.01)	0.824 (0.01)
ROP	LDA	RegAll	TS+GA+BW	72	0.742 (0.01)	0.213 (0.00)	0.812 (0.02)	0.736 (0.01)	0.336 (0.01)	0.247 (0.01)	0.839 (0.01)
ROP	QDA	IrregAll	SC+GA+BW	18	0.731 (0.01)	0.202 (0.00)	0.801 (0.02)	0.726 (0.01)	0.321 (0.01)	0.251 (0.01)	0.834 (0.01)
ROP	DT (*)	RegAll	SC+GA+BW	72	0.825 (0.00)	0.191 (0.01)	0.370 (0.02)	0.864 (0.00)	0.249 (0.01)	0.155 (0.01)	0.617 (0.01)
ROP	RF	IrregExcl6h	TS+GA+BW	72	0.817 (0.00)	0.259 (0.01)	0.688 (0.02)	0.828 (0.00)	**0.374** (0.01)	0.234 (0.01)	0.843 (0.01)
ROP	LR	IrregExcl6h	TS+GA+BW	72	0.763 (0.00)	0.220 (0.01)	0.759 (0.02)	0.763 (0.01)	0.339 (0.01)	0.251 (0.01)	0.842 (0.01)
ROP	SVM (*)	RegExcl6h	TS+GA+BW	72	0.919 (0.00)	0.813 (0.05)	0.006 (0.00)	0.998 (0.00)	0.012 (0.01)	0.257 (0.01)	0.833 (0.01)
ROP	*k*-NN	RegAll	TS+GA+BW	72	0.759 (0.01)	0.213 (0.01)	0.731 (0.02)	0.762 (0.01)	0.329 (0.01)	0.212 (0.01)	0.804 (0.01)

SE = standard error

Table A.2: Highest classification results complication- and classifier-specifically in terms of AUPR values

Complication	Classifier	Pre-processing	Feature selection	Monitoring time (h)	Accuracy (SE)	Precision (SE)	Sensitivity (SE)	Specificity (SE)	F_1 score (SE)	AUPR (SE)	AUROC (SE)
Mortality	GP_{linear}	IrregAll	TS+GA+BW	72	0.945 (0.00)	0.756 (0.04)	0.235 (0.02)	0.994 (0.00)	0.336 (0.02)	0.436 (0.02)	0.906 (0.01)
Mortality	GP_{m32}	IrregAll	ALL	72	0.942 (0.00)	0.664 (0.04)	0.247 (0.02)	0.990 (0.00)	0.336 (0.03)	0.435 (0.02)	0.918 (0.01)
Mortality	GP_{m52}	IrregAll	ALL	72	0.942 (0.00)	0.665 (0.04)	0.245 (0.02)	0.990 (0.00)	0.335 (0.02)	**0.437** (0.02)	0.918 (0.01)
Mortality	GP_{RBF}	IrregAll	ALL	72	0.942 (0.00)	0.661 (0.04)	0.249 (0.02)	0.990 (0.00)	0.337 (0.02)	0.436 (0.02)	0.920 (0.01)
Mortality	NB	RegExcl6h	SC+GA+BW	18	0.805 (0.01)	0.247 (0.01)	0.905 (0.01)	0.798 (0.01)	0.386 (0.01)	0.399 (0.02)	0.915 (0.01)
Mortality	LDA	IrregAll	TS+GA+BW	72	0.815 (0.00)	0.232 (0.01)	0.795 (0.02)	0.816 (0.00)	0.358 (0.01)	0.401 (0.02)	0.891 (0.01)
Mortality	QDA	IrregExcl6h	SC+GA+BW	18	0.822 (0.00)	0.259 (0.01)	0.912 (0.01)	0.815 (0.00)	0.402 (0.01)	0.409 (0.02)	0.918 (0.01)
Mortality	DT	IrregAll	ALL	36	0.908 (0.00)	0.345 (0.02)	0.418 (0.02)	0.942 (0.00)	0.370 (0.02)	0.246 (0.02)	0.683 (0.01)
Mortality	RF	IrregAll	ALL	36	0.905 (0.00)	0.379 (0.02)	0.663 (0.02)	0.922 (0.00)	0.475 (0.02)	0.420 (0.02)	0.922 (0.01)
Mortality	LR	IrregAll	TS+GA+BW	72	0.824 (0.00)	0.248 (0.01)	0.829 (0.02)	0.824 (0.00)	0.380 (0.01)	0.406 (0.02)	0.893 (0.01)
Mortality	SVM	IrregExcl6h	SC+GA+BW	36	0.933 (0.00)	0.492 (0.05)	0.131 (0.01)	0.987 (0.00)	0.189 (0.02)	0.404 (0.02)	0.916 (0.01)
Mortality	k-NN	IrregExcl6h	ALL	36	0.862 (0.00)	0.303 (0.01)	0.839 (0.02)	0.864 (0.00)	0.443 (0.01)	0.382 (0.02)	0.894 (0.01)
BPD	GP_{linear} (*)	RegAll	TS+GA+BW	72	0.784 (0.00)	0.655 (0.01)	0.519 (0.01)	0.889 (0.00)	0.575 (0.01)	0.602 (0.01)	0.853 (0.00)
BPD	GP_{m32} (*)	RegExcl6h	ALL	72	0.823 (0.00)	0.696 (0.01)	0.680 (0.01)	0.879 (0.00)	0.686 (0.01)	**0.715** (0.01)	0.888 (0.00)
BPD	GP_{m52}	RegAll	ALL	72	0.822 (0.00)	0.697 (0.01)	0.675 (0.01)	0.881 (0.00)	0.683 (0.01)	0.714 (0.01)	0.888 (0.00)
BPD	GP_{RBF}	RegAll	ALL	72	0.822 (0.00)	0.693 (0.01)	0.680 (0.01)	0.878 (0.00)	0.684 (0.01)	0.711 (0.01)	0.888 (0.00)
BPD	NB	RegAll	TS+GA+BW	72	0.763 (0.01)	0.562 (0.01)	0.795 (0.01)	0.750 (0.01)	0.657 (0.01)	0.594 (0.01)	0.844 (0.00)
BPD	LDA	RegExcl6h	TS+GA+BW	72	0.781 (0.00)	0.587 (0.01)	0.798 (0.01)	0.775 (0.00)	0.675 (0.01)	0.602 (0.01)	0.854 (0.00)
BPD	QDA	RegExcl6h	TS+GA+BW	72	0.792 (0.00)	0.608 (0.01)	0.769 (0.01)	0.801 (0.01)	0.678 (0.01)	0.617 (0.01)	0.846 (0.00)
BPD	DT	IrregExcl6h	SC+GA+BW	36	0.732 (0.00)	0.523 (0.01)	0.654 (0.01)	0.762 (0.01)	0.579 (0.01)	0.479 (0.01)	0.742 (0.01)
BPD	RF	RegExcl6h	ALL	72	0.795 (0.00)	0.606 (0.01)	0.811 (0.01)	0.788 (0.01)	0.693 (0.01)	0.700 (0.01)	0.883 (0.00)
BPD	LR	RegAll	TS+GA+BW	72	0.784 (0.00)	0.592 (0.01)	0.793 (0.01)	0.780 (0.01)	0.676 (0.01)	0.597 (0.01)	0.852 (0.00)
BPD	SVM (*)	RegAll	TS+GA+BW	72	0.779 (0.00)	0.647 (0.01)	0.500 (0.01)	0.890 (0.00)	0.561 (0.01)	0.592 (0.01)	0.851 (0.00)
BPD	k-NN	RegExcl6h	ALL	72	0.769 (0.00)	0.564 (0.01)	0.860 (0.01)	0.733 (0.01)	0.680 (0.01)	0.665 (0.01)	0.869 (0.00)
NEC	GP_{linear}	RegAll	ALL	72	0.968 (0.00)	1.000 (0.00)	0.000 (0.00)	1.000 (0.00)	0.000 (0.00)	0.124 (0.01)	0.777 (0.01)
NEC	GP_{m32}	RegAll	ALL	72	0.968 (0.00)	1.000 (0.00)	0.000 (0.00)	1.000 (0.00)	0.000 (0.00)	0.125 (0.01)	0.781 (0.01)
NEC	GP_{m52}	RegAll	ALL	72	0.968 (0.00)	1.000 (0.00)	0.000 (0.00)	1.000 (0.00)	0.000 (0.00)	0.127 (0.01)	0.780 (0.01)
NEC	GP_{RBF}	RegAll	ALL	72	0.968 (0.00)	1.000 (0.00)	0.000 (0.00)	1.000 (0.00)	0.000 (0.00)	0.125 (0.01)	0.778 (0.01)
NEC	NB	RegAll	TS+GA+BW	72	0.665 (0.01)	0.065 (0.00)	0.708 (0.03)	0.664 (0.01)	0.120 (0.00)	0.124 (0.01)	0.774 (0.01)
NEC	LDA	RegAll	TS+GA+BW	72	0.771 (0.00)	0.091 (0.00)	0.680 (0.03)	0.774 (0.00)	0.159 (0.01)	**0.142** (0.01)	0.773 (0.02)
NEC	QDA	RegAll	TS+GA+BW	72	0.859 (0.00)	0.103 (0.01)	0.449 (0.03)	0.873 (0.00)	0.166 (0.01)	0.113 (0.01)	0.747 (0.02)
NEC	DT (‡)	IrregAll	SC+GA+BW	48	0.920 (0.00)	0.139 (0.01)	0.296 (0.03)	0.940 (0.00)	0.184 (0.02)	0.094 (0.01)	0.618 (0.02)
NEC	DT (‡)	IrregAll	SC+GA+BW	72	0.920 (0.00)	0.139 (0.01)	0.296 (0.03)	0.940 (0.00)	0.184 (0.02)	0.094 (0.01)	0.618 (0.02)
NEC	RF	RegAll	ALL	72	0.936 (0.00)	0.222 (0.03)	0.303 (0.03)	0.957 (0.00)	0.232 (0.02)	0.134 (0.01)	0.778 (0.01)
NEC	LR	RegAll	TS+GA+BW	72	0.748 (0.00)	0.085 (0.00)	0.694 (0.03)	0.749 (0.01)	0.151 (0.01)	0.118 (0.01)	0.781 (0.01)
NEC	SVM	RegAll	ALL	72	0.968 (0.00)	1.000 (0.00)	0.000 (0.00)	1.000 (0.00)	0.000 (0.00)	0.124 (0.01)	0.782 (0.01)
NEC	k-NN	IrregAll	ALL	72	0.824 (0.00)	0.105 (0.01)	0.591 (0.03)	0.832 (0.00)	0.177 (0.01)	0.107 (0.01)	0.722 (0.02)
ROP	GP_{linear}	RegExcl6h	TS+GA+BW	72	0.919 (0.00)	0.682 (0.06)	0.030 (0.01)	0.996 (0.00)	0.051 (0.01)	0.252 (0.01)	0.838 (0.01)
ROP	GP_{m32}	RegAll	SC+GA+BW	72	0.920 (0.00)	1.000 (0.00)	0.000 (0.00)	1.000 (0.00)	0.000 (0.00)	0.260 (0.01)	0.842 (0.01)
ROP	GP_{m52}	RegAll	SC+GA+BW	72	0.920 (0.00)	1.000 (0.00)	0.000 (0.00)	1.000 (0.00)	0.000 (0.00)	0.260 (0.01)	0.841 (0.01)
ROP	GP_{RBF}	RegExcl6h	ALL	12	0.920 (0.00)	1.000 (0.00)	0.000 (0.00)	1.000 (0.00)	0.000 (0.00)	0.257 (0.01)	0.835 (0.01)
ROP	NB	RegAll	ALL	12	0.730 (0.01)	0.199 (0.01)	0.780 (0.02)	0.726 (0.01)	0.316 (0.01)	**0.262** (0.01)	0.827 (0.01)
ROP	LDA	RegExcl6h	TS+GA+BW	72	0.742 (0.01)	0.213 (0.01)	0.798 (0.02)	0.738 (0.01)	0.335 (0.01)	0.249 (0.01)	0.836 (0.01)
ROP	QDA	IrregExcl6h	SC+GA+BW	18	0.734 (0.01)	0.201 (0.01)	0.786 (0.02)	0.730 (0.01)	0.319 (0.01)	0.260 (0.01)	0.833 (0.01)
ROP	DT	IrregAll	ALL	36	0.855 (0.00)	0.213 (0.01)	0.306 (0.02)	0.902 (0.00)	0.247 (0.01)	0.161 (0.01)	0.607 (0.01)
ROP	RF	RegExcl6h	ALL	72	0.820 (0.00)	0.260 (0.01)	0.659 (0.02)	0.834 (0.00)	0.371 (0.01)	0.261 (0.01)	0.851 (0.01)
ROP	LR	RegAll	TS+GA+BW	72	0.755 (0.00)	0.213 (0.01)	0.761 (0.02)	0.755 (0.00)	0.332 (0.01)	0.255 (0.01)	0.841 (0.01)
ROP	SVM	RegExcl6h	TS+GA+BW	72	0.919 (0.00)	0.813 (0.05)	0.006 (0.00)	0.998 (0.00)	0.012 (0.01)	0.257 (0.01)	0.833 (0.01)
ROP	k-NN	IrregAll	TS+GA+BW	72	0.756 (0.00)	0.211 (0.00)	0.751 (0.02)	0.756 (0.01)	0.328 (0.01)	0.222 (0.01)	0.819 (0.01)

SE = standard error

Table A.3: Highest classification results complication- and classifier-specifically in terms of AUROC values

Complication	Classifier	Pre-processing	Feature selection	Monitoring time (h)	Accuracy (SE)	Precision (SE)	Sensitivity (SE)	Specificity (SE)	F_1 score (SE)	AUPR (SE)	AUROC (SE)
Mortality	GP_{linear}	IrregAll	SC+GA+BW	18	0.933 (0.00)	0.472 (0.05)	0.124 (0.02)	0.989 (0.00)	0.174 (0.02)	0.403 (0.01)	0.923 (0.00)
Mortality	GP_{m32}	IrregAll	SC+GA+BW	18	0.936 (0.00)	0.571 (0.04)	0.195 (0.02)	0.987 (0.00)	0.271 (0.02)	0.417 (0.01)	**0.928** (0.00)
Mortality	GP_{m52}	IrregAll	SC+GA+BW	18	0.937 (0.00)	0.576 (0.04)	0.198 (0.02)	0.988 (0.00)	0.274 (0.02)	0.418 (0.01)	**0.928** (0.00)
Mortality	GP_{RBF} (*)	RegAll	SC+GA+BW	18	0.936 (0.00)	0.592 (0.04)	0.215 (0.02)	0.987 (0.00)	0.291 (0.02)	0.427 (0.02)	0.927 (0.01)
Mortality	NB	IrregAll	SC+GA+BW	18	0.811 (0.00)	0.249 (0.01)	0.906 (0.01)	0.804 (0.01)	0.388 (0.01)	0.393 (0.01)	0.918 (0.00)
Mortality	LDA (*)	IrregAll	SC+GA+BW	18	0.804 (0.00)	0.243 (0.00)	0.912 (0.01)	0.797 (0.00)	0.382 (0.01)	0.387 (0.01)	0.919 (0.00)
Mortality	QDA	IrregAll	SC+GA+BW	18	0.822 (0.00)	0.262 (0.01)	0.910 (0.01)	0.816 (0.00)	0.404 (0.01)	0.390 (0.01)	0.919 (0.00)
Mortality	DT	RegExcl6h	SC+GA+BW	18	0.882 (0.00)	0.296 (0.01)	0.521 (0.02)	0.908 (0.00)	0.370 (0.01)	0.245 (0.01)	0.721 (0.01)
Mortality	RF (*)	IrregAll	ALL	36	0.905 (0.00)	0.379 (0.02)	0.663 (0.02)	0.922 (0.00)	0.475 (0.02)	0.420 (0.02)	0.922 (0.01)
Mortality	LR	IrregAll	SC+GA+BW	18	0.839 (0.00)	0.283 (0.01)	0.900 (0.01)	0.835 (0.01)	0.427 (0.01)	0.397 (0.01)	0.922 (0.00)
Mortality	SVM	IrregAll	SC+GA+BW	18	0.934 (0.00)	0.512 (0.04)	0.162 (0.02)	0.987 (0.00)	0.223 (0.02)	0.400 (0.01)	0.920 (0.00)
Mortality	k-NN	IrregExcl6h	ALL	36	0.862 (0.00)	0.303 (0.01)	0.839 (0.02)	0.864 (0.00)	0.443 (0.01)	0.382 (0.02)	0.894 (0.01)
BPD	GP_{linear}	IrregExcl6h	ALL	72	0.785 (0.00)	0.651 (0.01)	0.520 (0.01)	0.890 (0.00)	0.575 (0.01)	0.595 (0.01)	0.856 (0.00)
BPD	GP_{m32} (*)	RegExcl6h	ALL	72	0.823 (0.00)	0.696 (0.01)	0.680 (0.01)	0.879 (0.00)	0.686 (0.01)	0.715 (0.01)	**0.888** (0.00)
BPD	GP_{m52} (†)	RegAll	ALL	72	0.822 (0.00)	0.697 (0.01)	0.675 (0.01)	0.881 (0.00)	0.683 (0.01)	0.714 (0.01)	**0.888** (0.00)
BPD	GP_{RBF}	RegAll	ALL	72	0.822 (0.00)	0.693 (0.01)	0.680 (0.01)	0.878 (0.00)	0.684 (0.01)	0.711 (0.01)	**0.888** (0.00)
BPD	NB (*)	IrregAll	ALL	72	0.781 (0.00)	0.583 (0.01)	0.806 (0.01)	0.772 (0.01)	0.675 (0.01)	0.580 (0.01)	0.848 (0.00)
BPD	LDA (*)	IrregExcl6h	ALL	72	0.785 (0.00)	0.591 (0.01)	0.789 (0.01)	0.784 (0.00)	0.675 (0.01)	0.594 (0.01)	0.856 (0.00)
BPD	QDA	RegExcl6h	TS+GA+BW	72	0.792 (0.00)	0.608 (0.01)	0.769 (0.01)	0.801 (0.01)	0.678 (0.01)	0.617 (0.01)	0.846 (0.00)
BPD	DT	RegExcl6h	SC+GA+BW	72	0.728 (0.01)	0.521 (0.01)	0.669 (0.01)	0.752 (0.01)	0.584 (0.01)	0.473 (0.01)	0.748 (0.01)
BPD	RF (*)	RegAll	ALL	72	0.796 (0.00)	0.608 (0.01)	0.811 (0.01)	0.790 (0.01)	0.694 (0.01)	0.699 (0.01)	0.883 (0.00)
BPD	LR (*)	IrregExcl6h	ALL	72	0.792 (0.00)	0.601 (0.01)	0.798 (0.01)	0.790 (0.00)	0.684 (0.01)	0.588 (0.01)	0.856 (0.00)
BPD	SVM	IrregExcl6h	ALL	72	0.783 (0.00)	0.654 (0.01)	0.496 (0.01)	0.896 (0.00)	0.561 (0.01)	0.584 (0.01)	0.858 (0.00)
BPD	k-NN (*)	RegExcl6h	ALL	72	0.769 (0.00)	0.564 (0.01)	0.860 (0.01)	0.733 (0.01)	0.680 (0.01)	0.665 (0.01)	0.869 (0.00)
NEC	GP_{linear}	IrregAll	TS+GA+BW	72	0.968 (0.00)	1.000 (0.00)	0.000 (0.00)	1.000 (0.00)	0.000 (0.00)	0.113 (0.01)	0.793 (0.01)
NEC	GP_{m32}	RegExcl6h	TS+GA+BW	72	0.969 (0.00)	1.000 (0.00)	0.000 (0.00)	1.000 (0.00)	0.000 (0.00)	0.114 (0.01)	0.794 (0.01)
NEC	GP_{m52}	RegExcl6h	TS+GA+BW	72	0.969 (0.00)	1.000 (0.00)	0.000 (0.00)	1.000 (0.00)	0.000 (0.00)	0.113 (0.01)	0.793 (0.01)
NEC	GP_{RBF}	RegExcl6h	TS+GA+BW	72	0.969 (0.00)	1.000 (0.00)	0.000 (0.00)	1.000 (0.00)	0.000 (0.00)	0.114 (0.01)	0.790 (0.01)
NEC	NB	IrregAll	TS+GA+BW	72	0.671 (0.01)	0.066 (0.00)	0.712 (0.03)	0.670 (0.01)	0.121 (0.00)	0.106 (0.01)	0.785 (0.01)
NEC	LDA	RegExcl6h	TS+GA+BW	72	0.783 (0.00)	0.090 (0.00)	0.665 (0.03)	0.786 (0.00)	0.159 (0.01)	0.130 (0.01)	0.784 (0.02)
NEC	QDA	RegAll	TS+GA+BW	72	0.859 (0.00)	0.103 (0.01)	0.449 (0.03)	0.873 (0.00)	0.166 (0.01)	0.113 (0.01)	0.747 (0.02)
NEC	DT (‡)	IrregAll	SC+GA+BW	48	0.920 (0.00)	0.139 (0.01)	0.296 (0.03)	0.940 (0.00)	0.184 (0.02)	0.094 (0.01)	0.618 (0.02)
NEC	DT (‡)	IrregAll	SC+GA+BW	72	0.920 (0.00)	0.139 (0.01)	0.296 (0.03)	0.940 (0.00)	0.184 (0.02)	0.094 (0.01)	0.618 (0.02)
NEC	RF	RegAll	TS+GA+BW	48	0.933 (0.00)	0.173 (0.02)	0.275 (0.03)	0.954 (0.00)	0.205 (0.02)	0.130 (0.01)	0.802 (0.01)
NEC	LR	IrregAll	TS+GA+BW	72	0.740 (0.00)	0.079 (0.00)	0.671 (0.03)	0.742 (0.01)	0.140 (0.01)	0.103 (0.01)	0.789 (0.01)
NEC	SVM	IrregExcl6h	TS+GA+BW	72	0.969 (0.00)	0.984 (0.02)	0.000 (0.00)	1.000 (0.00)	0.000 (0.00)	0.106 (0.01)	**0.807** (0.01)
NEC	k-NN	IrregAll	ALL	72	0.824 (0.00)	0.105 (0.01)	0.591 (0.03)	0.832 (0.00)	0.177 (0.01)	0.107 (0.01)	0.722 (0.02)
ROP	GP_{linear} (*)	IrregExcl6h	TS+GA+BW	72	0.919 (0.00)	0.641 (0.06)	0.031 (0.01)	0.995 (0.00)	0.053 (0.01)	0.247 (0.01)	0.842 (0.01)
ROP	GP_{m32}	RegAll	ALL	72	0.920 (0.00)	1.000 (0.00)	0.000 (0.00)	1.000 (0.00)	0.000 (0.00)	0.243 (0.01)	0.846 (0.01)
ROP	GP_{m52} (†)	RegAll	ALL	72	0.920 (0.00)	1.000 (0.00)	0.000 (0.00)	1.000 (0.00)	0.000 (0.00)	0.243 (0.01)	0.846 (0.01)
ROP	GP_{RBF}	RegAll	ALL	72	0.920 (0.00)	1.000 (0.00)	0.000 (0.00)	1.000 (0.00)	0.000 (0.00)	0.245 (0.01)	0.846 (0.01)
ROP	NB (‡)	IrregAll	SC+GA+BW	48	0.728 (0.01)	0.203 (0.00)	0.821 (0.02)	0.720 (0.01)	0.324 (0.01)	0.231 (0.01)	0.837 (0.01)
ROP	NB (‡)	IrregAll	SC+GA+BW	72	0.728 (0.01)	0.203 (0.00)	0.821 (0.02)	0.720 (0.01)	0.324 (0.01)	0.231 (0.01)	0.837 (0.01)
ROP	LDA	IrregAll	TS+GA+BW	72	0.743 (0.00)	0.209 (0.00)	0.799 (0.02)	0.738 (0.01)	0.330 (0.01)	0.240 (0.01)	0.840 (0.01)
ROP	QDA	IrregAll	SC+GA+BW	18	0.731 (0.01)	0.202 (0.00)	0.801 (0.02)	0.726 (0.01)	0.321 (0.01)	0.251 (0.01)	0.834 (0.01)
ROP	DT	RegAll	SC+GA+BW	72	0.825 (0.00)	0.191 (0.01)	0.370 (0.02)	0.864 (0.00)	0.249 (0.01)	0.155 (0.01)	0.617 (0.01)
ROP	RF	RegExcl6h	ALL	72	0.820 (0.00)	0.260 (0.01)	0.659 (0.02)	0.834 (0.00)	0.371 (0.01)	0.261 (0.01)	**0.851** (0.01)
ROP	LR	IrregAll	TS+GA+BW	72	0.758 (0.00)	0.213 (0.00)	0.755 (0.02)	0.758 (0.00)	0.331 (0.01)	0.248 (0.01)	0.843 (0.01)
ROP	SVM	IrregAll	TS+GA+BW	72	0.919 (0.00)	0.797 (0.05)	0.003 (0.00)	0.998 (0.00)	0.006 (0.00)	0.246 (0.01)	0.836 (0.01)
ROP	k-NN	IrregAll	TS+GA+BW	72	0.756 (0.00)	0.211 (0.00)	0.751 (0.02)	0.756 (0.01)	0.328 (0.01)	0.222 (0.01)	0.819 (0.01)

SE = standard error