MASTER

Data mining in ECG data to predict patient deterioration in low resource settings

van der Putten, Jesse Jonathan

Award date:
2021

Technische Universiteit
Eindhoven
University of Technology

Department of Mathematics and Computer Science
Data Mining and Uncertainty in AI Group

Data mining in ECG data to predict patient deterioration in low resource settings

2IMC00 - Master Project

Jesse Jonathan van der Putten

Supervisor:
Prof. Milan Petkovic
Co-supervisor:
Dr. Roberto Rocchetta
Company supervisor:
Bart Bierling

Final version

Eindhoven, September 2021

Abstract

In this study we aim to develop a model that predicts if a hospital patient is deteriorating. The model can be applied to an affordable healthcare monitoring solution that tracks a patients vital signs and only electrocardiogram (ECG) data and basic patient-specific information will be needed for the model input. The model aims to predict if a patient is likely to die within the next 24 hours and is therefore most probably deteriorating.

To create and evaluate such a model we will be using a big amount of ECG waveform data. Peak detection is performed on ECG data and RR-intervals are extracted, from which statistical features on the intervals as well as the interval distribution are calculated to be used as input. A random forest is trained and evaluated using this data and further evaluation is performed on slight alterations of the model by including patient-specific variables and by applying sampling techniques to counter the class imbalance.

The results seem promising as the accuracy on deteriorating patients is high enough to catch most of them. Applying the model in a real world setting can potentially have a big positive impact on healthcare in low resource settings as the predictions can help allocate the limited resources better. Further research and real world testing are however needed to truly see the effect of such a model and to find out how the model can best be implemented given the needs in a specific healthcare setting.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In recent years data has been playing an increasingly important role in research and development. Machine learning (ML) and artificial intelligence (AI) are playing a more important role in our daily lives and are rapidly evolving in many different fields. One of the most crucial fields within this development, but which is slow in uptake and lacking behind, is healthcare. What makes this development so crucial is the ability to save lives, be it by assisting a doctor or fully taking over part of it's job [51]. Promising results have been show, but the implementation and adaptation of such improvements within healthcare take time. The reason for this is the need for rigorous testing and legal requirements needing to be satisfied, among other things [17], before moving the ML and AI assisted tools from research and development (R&D) to a real world setting.

This barrier has not slowed down the research and development efforts however. Experts have projected an AI in healthcare market of around $45B by 2026 from an estimated $4.9B in 2020 [2], around a 9x growth. This not only shows the massive growth that the healthcare space can still go through, but also the importance of money, both for purchasing and implementation, as well as for development. As a result the focus of many companies is on developed countries where a lot of money is available for healthcare and any improvements in it. Unfortunately this leaves the developing countries mostly on the sidelines, while these are the places that can benefit most from even small improvements in healthcare.

Basic monitoring systems and small technological improvement can already have a significant impact on the level of healthcare that can be provided in developing countries. This can as a result reduce mortality and morbidity in the general population, saving a great number of lives. The impact is greater in developing countries compared to developed countries as often deterioration is detected later in LRS leaving more room for improvement. A key aspect for such systems is affordability, as hospitals in low resource setting often lack the monitor systems to collect and process large amounts of data, as well as funds to purchase such devices. The focus on these low resource settings is thus to develop a light weight model that only takes a small amount of generally available variable as input and assists healthcare workers where possible.

The main goal of this thesis, is to develop a predictive model for healthcare in low resource settings, that can predict patient deterioration based on quantified ECG features and some patient specific inputs like gender, age and diagnosis. When talking about low resource settings we note both the limited monitoring systems and other hardware available at a hospital as well as the lack of staff and take both into account in our research. The main focus is therefore on providing information for healthcare workers on the well-being of their patients. More precisely the well-being of the patients that are currently being monitored by a monitoring system that has our model implemented while only using a limited amount of variables.

Chapter 2

Problem statement and research goals

2.1 Healthcare and the need for ML and AI

The COVID pandemic has shown us that there are still a lot of improvements to be made in the healthcare sector. Prediction, prevention, active care and recovery are all fields that can benefit from AI and ML enhancements. A positive note on this global pandemic is the willingness to invest more in research and development and speed up the implementation of new healthcare solutions. A major benefit that an algorithm has over a person is that it does not tire, is not effected by irrelevant circumstances and can deal with large amounts of data. On the flip side it can only use data that is measured or entered and can not make clinical observations itself. As the possibilities for AI grow we find more and more tasks that can be taken over by machines and algorithms. Think of smart robot assisted surgery [63] and virtual nurses [10] inside hospitals, but also of administrative tasks or auto generated reports.

An important problem of AI in healthcare that we wish to address in this work is real time tracking of vital signs for patients in the Intensive Care Unit (ICU). Vital signs themselves can give a good impression of the current well being of a patient but almost exclusively at the time of observation. Algorithms can in real time track the vital signs and detect changes that might be indicators of deterioration or a certain adverse event.

2.2 Healthcare in low resource settings

For various reasons ranging from funding to the availability of data from hospitals and the level of education, the focus on AI and ML developments in healthcare have been on developed countries. This leaves out most underdeveloped countries and regions that are already behind on healthcare innovations as is. This does mean that relatively small improvements, be it new or already in use technologies, can have a big positive impact on the state of healthcare in these regions. Overall healthcare in low resource settings can improve the most relatively to the further developed regions, but the lack of money and a solid base to build on makes it harder for new state of the art systems to be implemented. One of the main challenges, which we will focus on in this thesis, is to research and develop a ML model to support healthcare in low resource settings, e.g. by using a minimal amount of data.

2.3 Neonates

A big problem in developing countries is the relatively high death rate among newborns. One of the leading causes of death among babies born in developing countries is sepsis. What makes the death rate so much higher compared to other countries is often the delay in diagnosing of sepsis and as a result a delayed treatment. This can happen because of a lack of monitoring systems, but also because of a lack of staff, resulting in large time intervals between checks, both signs of how low resources alone can have a big negative impact. A real time monitoring system, that can be used to warn staff when a new born is showing signs of sepsis or overall deterioration can help prevent a large amount of these deaths. The main focus of this thesis was originally on this specific use case, but because of a lack of data, or access to it, researching this population was not possible.

2.4 Healthcare data availability

Not getting access to the neonatal healthcare data planned for this thesis pointed out another obstacle in ML research for healthcare. As healthcare data is highly sensitive personal data, it can be hard to obtain, both in a direct way where patient data is collected as well as in an indirect way where access to already collected healthcare data is required. A reason can be that the patient signed a consent form allowing only the entity that collects the data to be allowed to use it. One way to make healthcare data more publicly usable is by anonymizing it to the point there is it near impossible to connect a health record to any single person, effectively removing the sensitive personal side of the data. One such health data set is the Medical Information Mart for Intensive Care (MIMIC)-III database [44] [45], which has a separate MIMIC-III Waveform database [44] [57], hosted on PhysioNet [30]. The MIMIC-III database contains a wide range of healthcare data from over 40.000 patients, collected over an 11 year period between 2001 and 2012 at the Beth Israel Deaconess Medical Center in Boston. This data includes lab results, diagnosis, medication and billing information among many other variables. In this thesis we focus mostly on the age and diagnosis of a patient.

The MIMIC-III Waveform database contains raw waveform and numerical data from various monitoring devices like the electrocardiogram (ECG) from various leads, ambulatory blood pressure (ABP) and photoplethysmogram (PPG). A separate subset [56] is also available with subject IDs that are matched with the subject IDs from the MIMIC-III database. As we will be using data from both sets the matched subset will be used in this thesis.

The MIMIC-III database however lacks in data from minors age 18 and under, as only a handful can be found among all patients. A supplement was proposed by Zeng et al [91] which includes data from the Children's Hospital Zhejiang University School of Medicine's pediatric ICUs. The database is structured in a similar way to MIMIC-III, but there is no subsequent waveform data made available to complement the database.

2.5 Main goals

The main goal of this thesis is to create a model that predicts if a given patient is deteriorating based on its ECG data and some patient-specific information. In the list below the main goals and targets we will discuss in this thesis are summarized.

- Develop a ML model for healthcare that can be used and deployed in low-resource settings. Important aspects for this are low computational power of the hardware deployed in these hospitals. Also limited need for input data e.g. only using one vital sign and basic patient-specific information instead of a vast amount of data like multiple lab results, patient background and multiple vital signs data streams, can make the model more widely applicable.

- Train and evaluate ML models to predict patient deterioration using only an ECG signal for input. From this ECG signal features will be extracted and used as model input.

- Optimize the model and tune its performance using a small amount patient-specific data, e.g. age and gender.

- Further optimize the model and tune its performance using patient-specific diagnosis information that could have a big effect on the ECG data, i.e. heart related diagnoses.
- Gain knowledge on feature importance for each classification.

Chapter 3

Literature review

3.1 Healthcare in low resource settings

In a 2009 paper, Naicker et al. [62] compared healthcare numbers of developing countries like Africa with the rest of the world. The severe shortage of healthcare workers becomes immediately apparent when looking at the amount of healthcare workers per 1000 inhabitants, which stands at only 2.3 in Africa versus 24.8 in America. When looking at just doctors, we see that most sub-Sahara countries have only 1 or 2 doctors per 10000 inhabitants. This is for a big part due to educated healthcare professionals leaving the country for better jobs abroad. Losing these professionals has a negative impact on healthcare in Africa, but also educating people to become healthcare professionals in the first place is a harder task in developing countries. Initial research done at the Medical Library, College of Medicine at the University of Nigeria also showed a lack of ICT infrastructure needed for information gathering and knowledge transfer [89]. Such an infrastructure is needed to bring healthcare education to a higher level as well as to support the research and development of healthcare solutions that require locally collected data.

No matter the reason, a clear image is shown of the lack of healthcare professionals in Africa. New innovations in healthcare technology can fill part of this gap and give some much needed assistance.

3.2 ML and AI in healthcare

Applying ML using big data in healthcare [9] was a logical next step after successfully using it in various other fields like finance and politics [61], [66]. With this comes also the shift to a more patient centered care [77]. A key factor is the use of electronic health records (EHRs) which hold all personal healthcare information of a patient, which has seen a lot of growth after the Health Information Technology for Economic and Clinical Health Act [11]. This data and the information gained from it has already supported the diagnosis procedure which previously was purely symptom based. The amount of data gathered in EHRs and in healthcare as a whole has grown exponentially over the past decade, which went hand in hand with a broader adoption of EHR usage [47]. The data collected in these health records are important to the development of new ML solutions. They can help with diagnosis and to optimise a patients hospital stay, but also find the optimal date for a check-up or risk assessment [41]. An EHR can hold a wide variety of data, including textual data which can be used in natural language processing [42], patient specific data like age and gender or data specific to a doctors visit or hospital stay like symptoms, diagnosis or vital signs if monitored.

3.3 Predicting deterioration

3.3.1 Early warning scores

There are multiple ways in witch the term deterioration in healthcare can be interpreted. This can for instance be predicting the chance of re-hospitalisation after discharge based on past information. In this thesis we will focus on real-time prediction of deterioration for patients in an ICU based on live vital sign monitoring. To predict deterioration the severity of the illness is important. The initial focus for this thesis was on neonates, for which multiple severity scores have been developed. The Clinical Risk Index for Babies (CRIB) is a severity score that predicts mortality for newborns with a gestational age of less than 32 weeks [14] and was developed in 1990 using logistic regression to find 6 variable that were most indicative to predict mortality. It was later improved in CRIB-II [64] by using fewer variable that most notably aren't effected by care given. Another score that was created around that time was the Score for Neonatal Acute Physiology (SNAP) [76]. This score is based on 28 variables collected within the first 24 hours after birth. This makes the score harder to compute as any missing values would effect it. An extension on this was made called SNAP-PE (Perinatal Extension), which included 4 variables [74]. Because of the difficulty of getting all these variable both the SNAP and SNAP-PE scores were updated to include only 6 and 4 variable respectively and are based on measures done in the first 12 hours after birth [75]. More scores based on patient data have been created, like the National Institute of Child Health and Human Development (NICHHD) score [39] and the Neonatal Mortality Prognosis Index (NMPI) [28]. One notably different score is the National Therapeutic Intervention Scoring System (NTISS) [32], as its variables are derived from the treatment given to a patient as opposed to measures from the patient itself.

These scores were still manually calculated based on the attained variables, but with the rapid development of computer technology and later ML, predictions and scores can be calculated automatically and in real time. One such scores is the Targeted Real-time Early Warning (TREW) Score [36] which predict the risk of patients developing sepsis. It uses both vital signs as well as regular lab results to predict if a patient is at risk of developing sepsis.

3.3.2 Real time tracking

Real time tracking of vital signs is used in various different ways to predict deterioration. Joshi et al. [46] used Heart Rate Variability (HRV) features, respiratory features and an estimated movement based of ECG signals of preterm infants to predict late-onset sepsis (LOS). The average HR acceleration response and the respiratory interdecile range were shown to be two good indicators of LOS, seeing a significant change multiple hours before onset. Using vital signs to track neonates is of extra importance as they are quite vulnerable, especially when born pre-term and/or underweight, as well as for the fact that they lack the ability to communicate clearly with a healthcare worker about their own well being. For this reason a lot of studies have been preformed on newborns tracking the heart rate[1], pulse oximetry monitoring[2], respiratory rate[3] and blood pressure[4].

3.3.3 ECG-based

Research even closer to the core of this thesis are those using only ECG data to make predictions. ECG data is used to monitor a patient's heart rate, but a lot more information can be extracted from it. A very informative variable that can be calculated from an ECG signal is the HRV, which measures the variability between consecutive heart rates and can be a good indicator for someones current health state as well as a predictor for possible deterioration [3]. Fairchild and Aschner

[1] [26] [34] [59] [80] [20] [4] [83] [82] [29] [27] [86] [33] [24] [25]

[2] [88] [16] [25] [6] [69] [19] [81] [72] [87] [18] [50] [12]

[3] [88] [21] [65] [55] [38] [84]

[4] [15] [31] [53] [7] [49] [37] [52] [68] [79] [78]

[23] combined the HR with sample asymmetry and sample entropy to calculate the Heart Rate Characteristics (HRC) [58]. Real time tracking of the HRC was used to monitor neonates and predict the onset of sepsis and to create a warning score that indicates the likelihood of a patient getting sepsis. A monitoring system called 'HeRO' was created for real world use of the predictive model. A randomized trial of 3003 neonates was conducted to research the effect of their early warning score on mortality. They concluded that the overall mortality reduced by over 20% when using the HeRO score [22].

3.3.4 MIMIC-III

Shifting the focus towards the MIMIC-III database and an adult population, we see that it has been used to predict deterioration [43] as well. In one such study, Hou et al. [40] created and tested three models to predict 30-day mortality for patients with sepsis. As input geographic data, vital signs and laboratory values were used of patients who stayed in the ICU for a period of at least 24 hours. The vital signs used were numeric vital signs stored in the MIMIC-III database as opposed to the raw vital sign data offered in the MIMIC-III Waveform database.

Taoum et al. [85] did use a combination of the MIMIC-II and its waveform databases. The aim of their study was to predict Acute Respiratory Distress Syndrome (ARDS) for patients in an ICU. They used HR and Breathing Rate (BR) to make their predictions. Their results have shown that the HR was more informative in making the prediction.

3.4 ECG features

An ECG displays the energy flow of the heart. The most visible and easy to spot part in an ECG chart is called the QRS complex, as shown in Figure 3.1[5], which displays the main spike of energy released during a heart beat. The time between heartbeats is called the RR-interval, as it is the time between R-peaks.

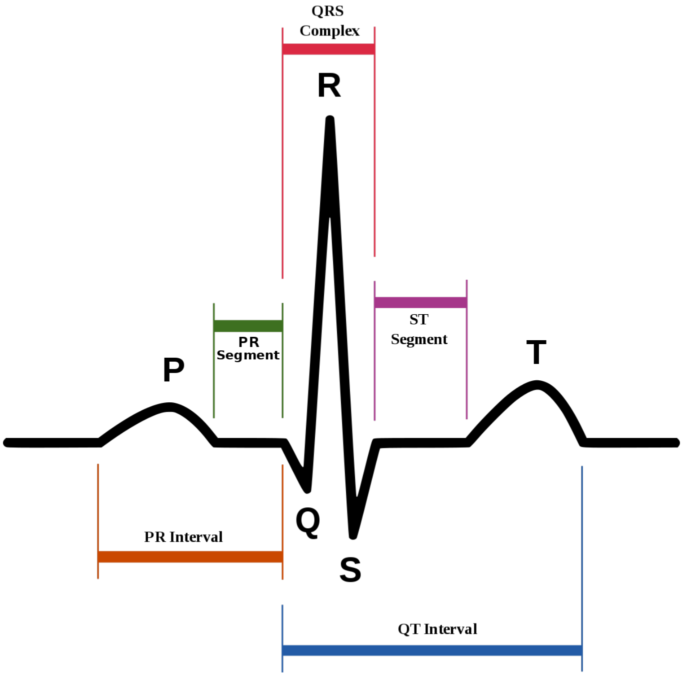

Figure 3.1: Typical QRS complex.

[5] https://en.wikipedia.org/wiki/QRS_complex#/media/File:SinusRhythmLabels.svg

Features extracted from an ECG signal can either be frequency based or time based. Because of the way the ECG data in the MIMIC Waveform database is scaled and processed it is unsuitable for certain frequency domain algorithms, as stated by the publishers. For this reason the ECG features used in our study are limited to time based features.

ECG data has been used to improve detection of deterioration when used on top of existing models [60] and in many studies ECG derived data has been used for early warnings of patient deterioration or mortality [71], but almost exclusively in combination with other data or on very specific diagnoses or conditions. HRV is often a key indicator and has been used to predict deterioration in patients with specific deceases, e.g. sepsis [8], chronic kidney decease [13] or sudden death in epilepsy [73]. For this reason the HRV measures are one of the key features to extract from our data.

HRC have been used in multiple studies on neonatal sepsis and is a strong indicator for predicting sepsis and sepsis-like illnesses. A key feature used in the HRC is the sample asymmetry (SampAsy), which has shown to be a good indicator by itself in the study by Joshi et al[46]. In this same study the interdecile range, meaning the difference between the 10th and the 90th percentile, of the respiratory rate was also used and shown to be a good indicator of patient deterioration. The interdecile range can also be used in the heart rate detected as a possible indicator, although no research has been found using this method.

A popular way of visualizing HRV is by creating a Poincaré plot, which shows the difference in successive RR-interval length, as shown in the example in Figure 3.2 where X_i is the ith RR-interval.

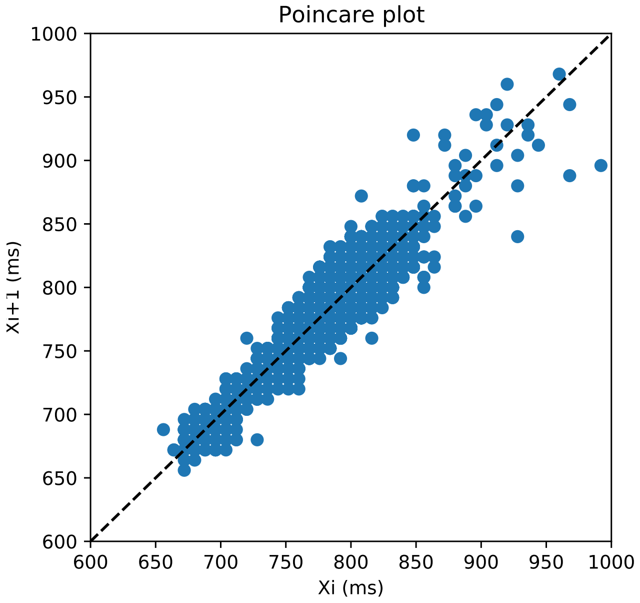

Figure 3.2: Example Poincaré plot from patient 822, hour 14

The plot shows the temporal correlations of RR-intervals within a certain time frame, e.g. one hour. The deviation along the diagonal shows the variability over the longer term, while the deviation perpendicular to the diagonal shows the variability over the short term between successive heart beats. The symmetry in the plot represents the symmetry of the heart beats, making asymmetry easy to spot. Guzik et al. [35] created an index (GI) that calculates the contribution of deceleration's to short-term HRV by calculating the ratio of the distance of point below the diagonal to the overall distance of the point in the plot.

Another way to assess the asymmetry in the RR-interval distribution is to calculated the ratio of the number of points below the diagonal versus the total amount of points, as introduced by Porta et al. [70].

3.5 Current gap in literature

What stands out when reading the latest work on predictive models for healthcare is that they tend to focus on a very specific subgroup such as a specific disease, e.g. sepsis, and/or a specific age group. From a research perspective this is interesting and highly accurate models are created as a result. From an application perspective this is however not always very helpful, as some models need a lot of input data to work, or are only applicable on a small subgroup of patients. In this thesis we try to tackle this problem by using few input data and training the model on a wide variety of patients. Another reason why new models might not perform well in a real world setting is the knowledge gap between healthcare and machine learning knowledge. Extensive machine learning knowledge is not needed to use a model, but by not knowing what a model bases its prediction on a healthcare worker might be skeptical or simply ignore the model outcome. In this thesis we will use a relatively easily interpretable model which enables the possibility to interpret and translate the model workings to support the prediction outcome.

Chapter 4

Methods

4.1 Data exploration and cleaning

The MIMIC-III database contains a vast array of different data, both patient specific as well as generic data like billing codes. Each patients has a unique *SUBJECT_ID* which is used among different tables in order to match entries for the same subject. Dates have been shifted in order to anonymise the data, but a patient's age can still be derived by calculating the difference between the date of birth and the admission date. For ages 89 and above the dates were further shifted to increase anonymity and are therefore excluded from this study. The initial focus when using a different database was to focus on neonatal and/or pediatric patients, but as seen in Figure 4.1 there are close to no patients under the age of 18 in the MIMIC database. For this reason a change in approach was needed and the population for our study were changed to adults, age 18 through 88. In total the database contains 46520 unique patients, of which 36559 remained after filtering by age.

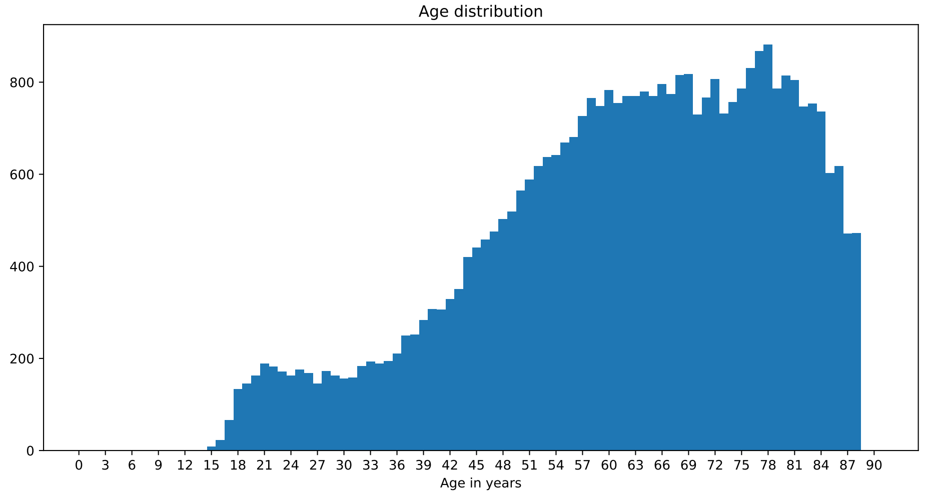

Figure 4.1: Age distribution of patients in the MIMIC-III database, excluding 89+

In the next step we looked at what patients had waveform data by matching the *SUBJECT_ID* from the admissions with those in the matched waveform subset. In the base directory of the waveform database a *RECORDS* file was kept that contains all folders in the database. The first layer of folders are named after the first 2 digits of a *SUBJECT_ID* appended to the letter 'p', with the folders in the second layer being named after the *SUBJECT_ID*, also appended to the letter 'p', e.g. *base/p01/p014837/*. To match the IDs we downloaded the *RECORDS* file, extracted the IDs from the folder names and matched these with the 36559 subjects left after filtering. This resulted in 9874 subjects age 18 through 88 that had waveform data available.

For our research we wanted to focus on the final 24 hours of an ICU stay. One of the tables in the MIMIC-III database named *ICUSTAYS* contains the length of stay variable, indicating in days how long the ICU stay of a patient was, varying from a couple minutes for some to multiple weeks for others. At this stage multiple ICU stays by the same patient were all included as long as they were at least 24 hours in length, resulting in 14474 ICU stays by 8961 unique patients.

The next step in filtering was done by checking if the waveform data contained the ECG lead II signal, a signal where peaks are clearly visible. In order to do this the header file of an ICU stay has to be read. Each ICU stay's waveform data is divided into segments and each segment has a header and a data file. The first line of the header file starts with the name of the segment, which is denoted as the waveform ID, a unique ID for each ICU stay, followed by a 4 digit segment number, e.g. *3544749_0001*. This is followed by the amount of channels recorded, e.g. 3, 2 ECG channels and 1 blood oxygen measurement, the signal frequency, which is 125Hz for all data, the amount of data points in the segment and ends with the time of the segment. The other lines in the header file give information on the different signal channels in the data file. An example header file can be found in the appendix section A.1. To retrieve the file name we had to download each patient's *RECORDS* file and extract all waveform IDs and segment numbers. Using this each ICU stay's data signals were checked by reading the header files and patients without ECG lead II (denoted by a 'II' at the end of the line) were excluded from the population. After this filtering step 8213 patients and 13591 unique stays were left.

Although the patients had been filtered by an ICU stay of at least 24 hours, this does not mean that there is 24 hours of ECG data available. At 125Hz the sum of segments has to be at least $125 * 60 * 60 * 24 = 10800000$ data points in size to contain 24 hours of data. To check this the segment length of each segment per unique stay was added after reading the first line of each header file. Filtering out another 1690 patients and 2770 stays that did not contain enough data, leaving 6523 patients and 10821 ICU stays.

For any patient with multiple ICU stays one can ask the question if they are related. Multiple stays can happen is a patient is moved to surgery in between, or had been discharged but gotten sick again and had to return for instance. Because of the uncertainty as to why a patient could have multiple ICU stays we decided to only look at the final stay. A patient couldn't have died during an earlier ICU stay and there is a chance that he or she wasn't fully or rightfully discharged so labeling these stays as discharged or expired would not be possible without any uncertainty. An earlier ICU stay can also have an effect on a later one, but as the outcome of the last ICU stay would still be discharged or expired, and final, having had earlier ICU stays would not effect the labeling of these stays. This brings the total stays equal to the amount of patients, namely 6523.

The ECG features will eventually be split into hourly segments for input to the classification model. To get an overview of the waveform data we created a dataset containing the *SUBJECT_ID*, waveform ID, segment number, start time, end time, hour (1-24) and gap length, in case there were gaps or 00:00 otherwise. This way we know for each hour in what data segment file we need to look and if there are any significant gaps between them. When looking at the last 24 hours of a stay we have to start at the end and work backwards in order to go through the data. The main challenge that arises here is that the segments are of variable length and that there may be gaps between segments. This can create a couple different scenarios. Let x be the length of a segment, r the remaining time in the hour we are gathering information on, g the length of a gap and h the length of one hour.

1. The segment is less than an hour long, $x < h$.

   In this case we have to keep track of how much of the hour we have processed before we move to the next segment. The time that is remaining in the hour is $r = h - x$, as shown in Figure 4.2.

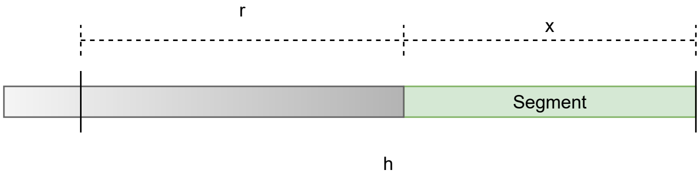

Figure 4.2: An illustration of a segment that is shorter than an hour

2. The segment is exactly one hour long, $x = h$.

   In this case we can use the entire segment as one hour of data.

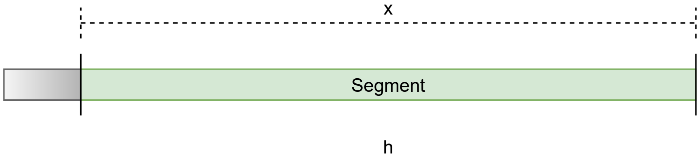

Figure 4.3: An illustration of a segment that is exactly one hour long

3. The segment is more than an hour long, $x > h$.

   In this case we check how many hours fit in the segment and split it up. We denote the amount of whole hours in the segment as n, where $n = \lfloor \frac{x}{h} \rfloor$ and what is left is denoted as l, where $l = x \mod h$. After processing an hour the length of x decreases by h, so that $x_{new} = x_{old} - h$. After processing n hours we are left with $x < h$ and x will be handled as in 1 (Figure 4.2), where $x = l$. In the cases that $l = 0$ this scenario can be split up into multiple cases of scenario 2.

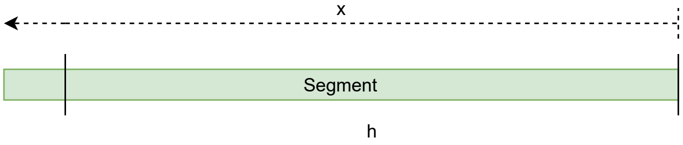

Figure 4.4: An illustration of a segment that is longer than an hour

The above scenarios work for the first segment or when starting an hour at the start of a segment, but when a part of an hour is left by the end of the segment, $r \neq 0$ and $l \neq 0$, we move to the next segment to check if it contains the remaining data to get to one hour. The next segment can again be of various lengths. The data already added to the hour we are processing is denoted by l. l can be what was leftover in the previous segment as in scenario 3 or the length the the previous segment as in scenario 1, which are equal in this case. With r data missing to fill the hour we can again fall into three different scenarios.

4. The segment is shorter than the time remaining in the hour, $x < r$.

   In this case the segment is added to the hour, $r_{new} = r_{old} - x$ and $l_{new} = l_{old} + x$ and we move on to the next segment.

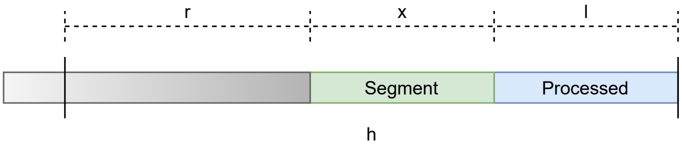

Figure 4.5: An illustration of a segment that is shorter than the remaining time

5. The segment is the exact length of the time remaining in the hour, $x = r$.

   Although unlikely, in this scenario we can use the segment to complete our hour and start a new hour on the next segment as in scenarios 1-3.

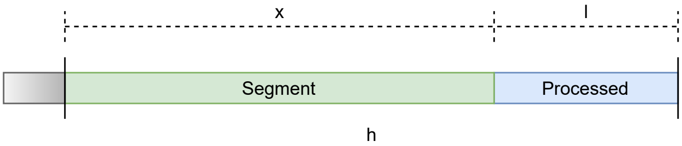

Figure 4.6: An illustration of a segment that is the same length as the remaining time

6. The segment is longer than the time remaining in the hour, $x > r$.

   In this case we add what is left to the current hour and move to the next. This time is subtracted from x so $x_{new} = x_{old} - r$. We can now treat the remaining x as in scenarios 1-3.

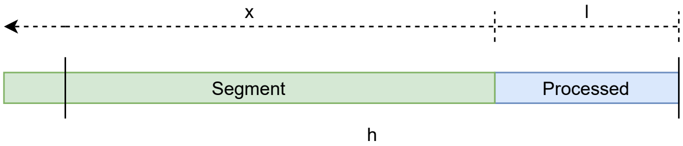

Figure 4.7: An illustration of a segment that is longer than the remaining time

Whenever we move to a new segment however, as after scenarios 1, 2, 4 and 5, there is a chance that there is a gap of missing data between them. To calculate this we calculate the end time of the new segment and compare it with the start time of the segment we just finished. As the time is expressed in 24 hours, there exists a chance that the last segment started just after midnight, say 00:05, and the new segment ended just before midnight, say 23:55 (note we are calculating backwards in time). In this case we cannot simply look at the difference in time, as it would return a gap of -23:50, were as the actual gap was only 10 minutes. To deal with such scenarios we added the total time in one day to the difference modulo the total time in one day. A gap can again have various lengths. The length of a gap is denoted as g and there are four different scenarios to consider.

7. The gap is shorter than the time remaining in the hour, $g < r$.

   In this case we note the gap in our dataset and add the length of the gap to the time we have processed in our hour, $l_{new} = l_{old} + g$ and subtract is from the remaining time, $r_{new} = r_{old} - g$. With what is left we go to the next segment and proceed as in scenarios 4-6.

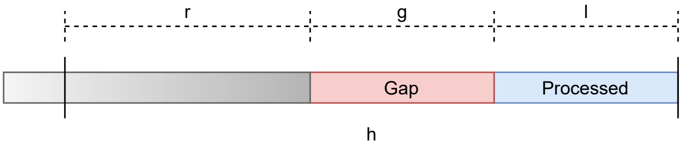

Figure 4.8: An illustration of a gap that is shorter than the remaining time

8. The gap is exactly the length of the time remaining in the hour, $g = r$.

   Although unlikely, in this scenario we can note the gap in our dataset and start the new hour at the start of the next segment, as in scenarios 1-3.

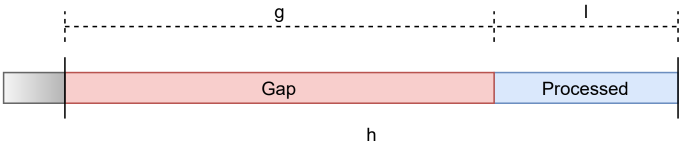

Figure 4.9: An illustration of a gap that is the same length as the remaining time

9. The gap is larger than the time remaining in the hour, $g > r$.

   In this case we can add the remaining time to our hour and note it as a gap in the dataset. The length of the gap is than decreased by the time we had remaining in the hour, $g_{new} = g_{old} - r$. The remaining length of the gap can be processed as in scenarios 1-3, with the difference being the data being processed being a gap instead of a segment, effectively replacing x by g but following the same procedures.

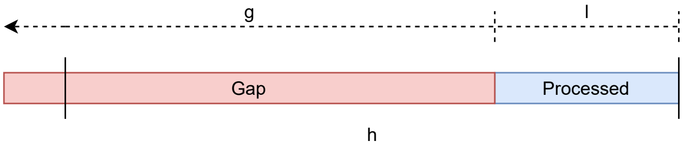

Figure 4.10: An illustration of a gap that is longer than the remaining time

10. A new hour starts with a gap.

    In the unlikely case of scenarios 2 and 5 there is a chance that the next segment doesn't align with the start of the new hour, meaning it will start with a gap. In this case we also handle the gap g as we do x in scenarios 1-3, with gap lengths being noted in the dataset.

In any case where an hour is fully processed to check if it was the 24th hour. If so we stop and move to the next patient.

In all cases where either $g > r$ or $x > r$ with, $r \leq h$, we split either g or x up into pieces smaller or equal to the data needed to process an hour, r. As a result, in each possible scenario we have $h = l + x + g + r$ where r is always filled up by either x or g as $x_{total} \geq 24 * h$ and thus no data is ever missed when processing the waveform data.

We decided that data of a patient will only be used if no more than 5% of data is missing in any single hour. Having too many gaps and the uncertainty of the exact length of a gap in days makes for a strict limit in allowance of missing data between segments. After filtering our patients with more than 5% of missing data between segments in any single hour we were left with 4353 patients with sufficient ECG data in the last 24 hours of their stay, totalling $4353 * 24 = 104472$ hours.

4.2 ECG feature extraction

All features that will be extracted from this data are time based and calculated from the RR-intervals. In order to get the RR-intervals we first need to detect the R-peaks. To open and view the data files the WFDB library [54] from MIT is needed. This library can read, process and plot the waveform files and will be used to find the R-peaks in the ECG data.

4.2.1 Peak detection

In the previous section we have analysed the waveform data header files and created a table which contains information on what hourly data can be find in which files. This information can now be used to determine what parts of a data file need to be read in order to get one hour of R-peaks. When multiple files contain data for one hour the newly found peaks will be shifted according to the time that has already been processed. this way each hour starts at 0 and ends after $125 * 60 * 60 = 450000$ data points. The WFDB library's PROCESSING.QRS.GQRS_DETECT() can be used to find peaks in the ECG signal. An example using 2000 data points, or 16 seconds, can be seen in Figure 4.11, where an X denotes a found peak.

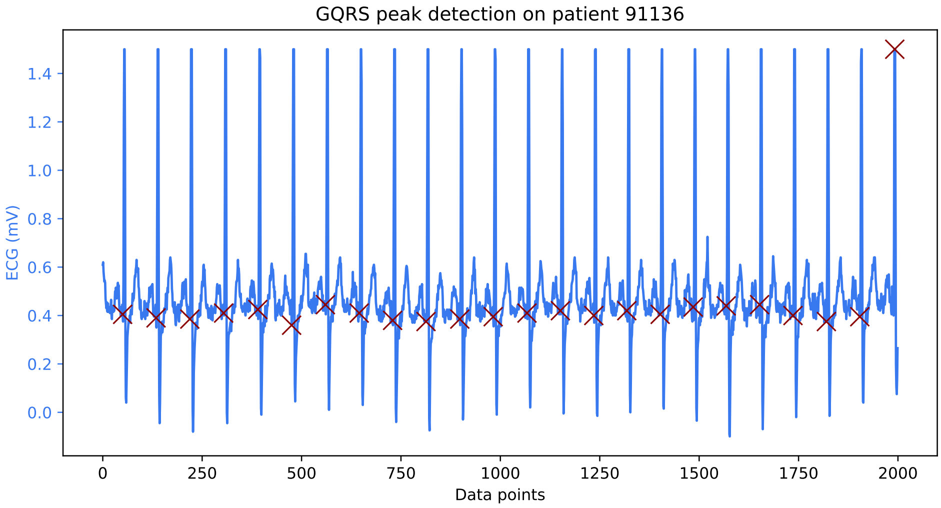

Figure 4.11: Initial detected peaks

Although the peak detection has returned peaks, they aren't the R-peaks we are looking for. To correct this PROCESSING.PEAKS.CORRECT_PEAKS() is used, indicating that we are looking for upwards pointing peaks. The result can be seen in Figure 4.12. This method returns the location of the found peaks. To get the RR-intervals we calculate the distance between peaks and save these per patient and per hour.

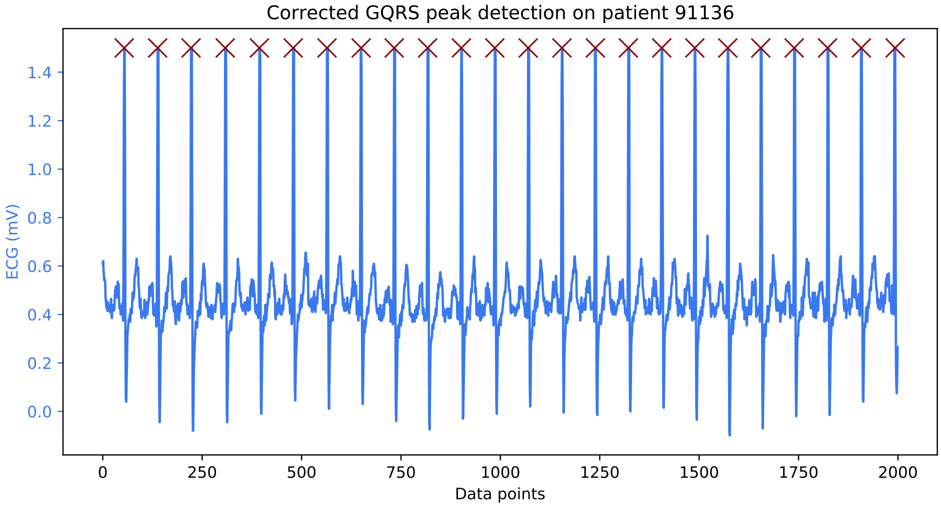

Figure 4.12: Corrected detected peaks

4.2.2 Outlier detection

Although the gaps between data segments are there to account for missing data, not all data files contain a clear ECG signal. As a result a lot of the hours calculated can still have more than the 5% threshold of missing data on which they were filtered earlier. Also corrupted files or gaps of data stored as Not a Number (NaN) values were found, and as no peaks could be detected in such files they were excluded from the dataset. In total 35939 hours of data were excluded after being unable to detect peaks due to missing values and 4281 hours because of the amount of peaks being below the threshold of an average HR of 30 over a time period of at least 55 minutes, namely $30 * 55 = 1650$ peaks. An upper limit of an average HR of 230 over a full hour was also set as an exclusion criteria, but no hour contained that many detected peaks. In total this leaves 64252 hours of data.

The peak detection and correction correctly return peaks in most cases, however, this process returns unexpected results when ECG data have poor quality. This can occur, for instance, because of poor skin preparation or a patient movement leading to a loss of electrode-to-patient contact. For this reason there can be outliers in the RR-intervals. To detect these outlier we calculate the mean and standard deviation for an hour of data and set the threshold for exclusion as any interval with a length more than 2 standard deviations from the mean. This method is known as the z-score, where the score of a point is calculated as

$$z = \frac{\|x - \mu\|}{\sigma} \tag{4.1}$$

x is our sample interval, μ our mean interval length and σ out standard deviation. A max distance of 2 standard deviations thus excludes intervals with a z-score of $z > 2$. As an example we look at a randomly selected patient and hour, patient 328 and hour 3. After the peak detection and RR-interval calculation we have a shortest interval of 296ms and a longest interval of 1592ms. These correspond to a HR of 203 and 38 respectfully, while the average HR was only 81 beats per minute. As such a swing in HR is extremely unlikely and that these extremes only occurred few times as they are barely visible in the distribution in Figure 4.13, we can safely say these are outliers that occurred due to either a faulty signal or peak detection.

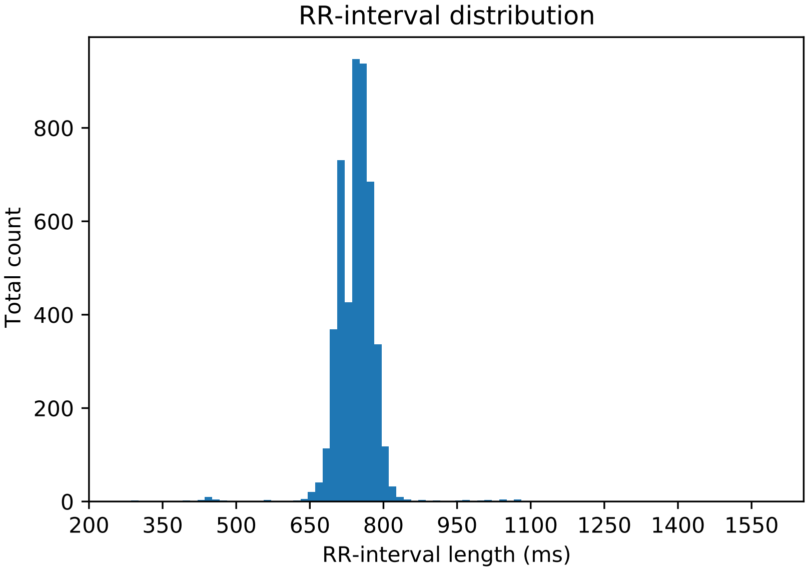

Figure 4.13: RR-interval distribution after peak detection

When applying our z-score and removing the outliers we find a shortest interval of 656ms and a longest interval of 832ms, corresponding to a HR of 91 and 72 respectfully. With the average still being 81 beats per minute this distribution seems more likely, as seen in Figure 4.14.

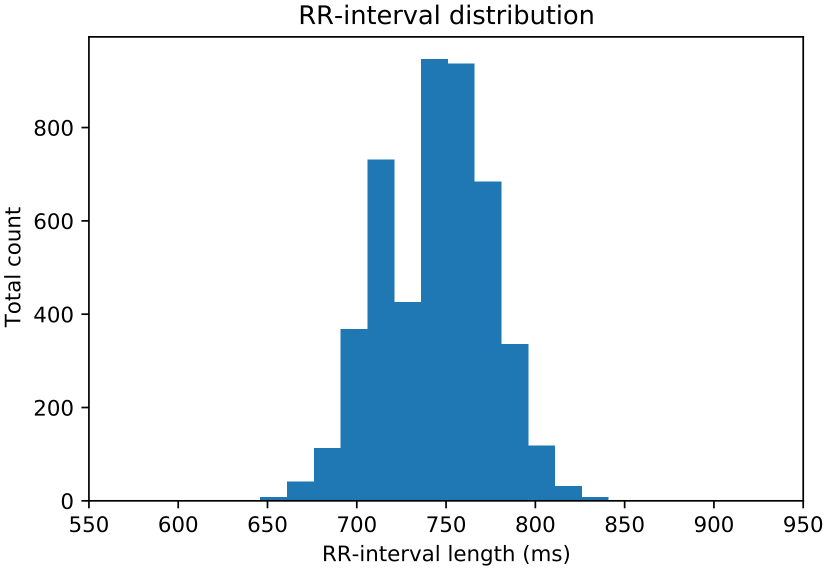

Figure 4.14: RR-interval distribution after outlier removal

An example where the HR is not (near) normal distributed is for a patient with atrial fibrillation (AF). AF is a condition where the HR goes up significantly for a short period of time before coming back down and is common among patients in ICU [5]. This condition is easy to spot as the RR-interval distribution shows 2 clear peaks, as shown in Figure 4.15. Most outliers are seen on the right side of the distribution, where the RR-intervals are the longest. Note that the x-axis ranges from the lowest to the highest RR-interval lengths, where some lengths only occur one to a few times, making them hard to see in the graph.

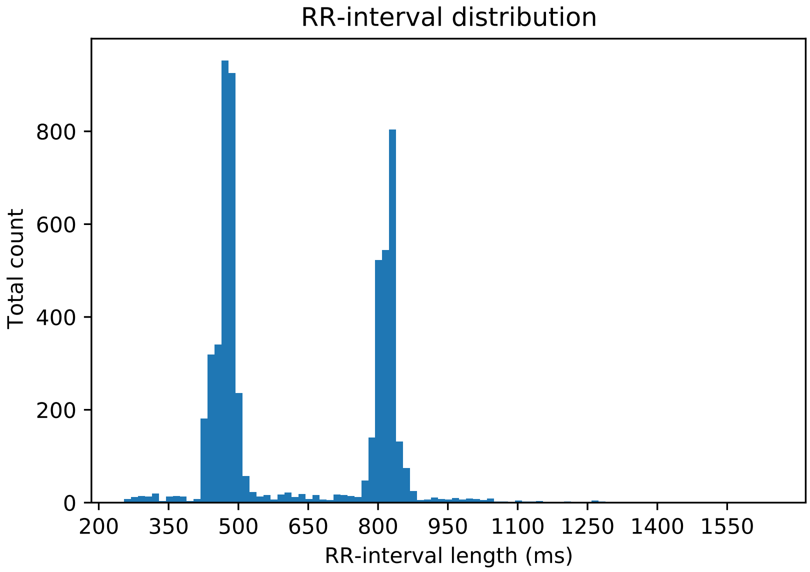

Figure 4.15: RR-interval distribution of a patient with AF after peak detection

When applying the same outlier detection with a z-score of 2, the longest interval was 984, down from 1640. It is visible though that with more extreme variability in intervals fewer outliers might be detected, as shown in Figure 4.16.

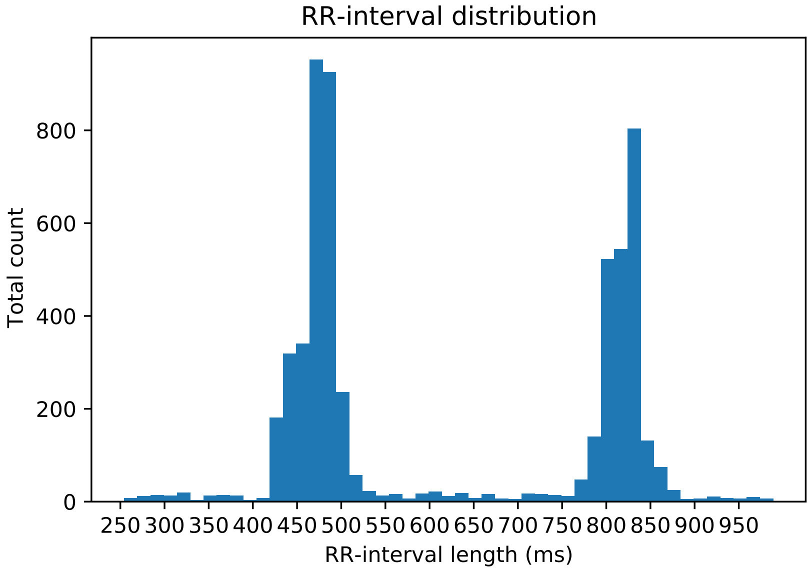

Figure 4.16: RR-interval distribution of a patient with AF after outlier removal

Although big changes in HR can occur in patients in the ICU, a significant amount of corresponding intervals is to be expected when this occurs, for instance when the patient is in pain or undergoes stress. A significant enough amount of such intervals would alter the mean and standard deviation, making them less likely to be classified as outliers. As the data is vast and includes patients of all ages and diagnoses it is out of the scope of our research to specifically pick outliers as some might be patient specific physiological changes while others are simply faulty data. For this reason the minimum and maximum intervals will not be used as inputs for the model, whereas they could be a telling variable in other cases. Also the z-score of 2 is used to not exclude or include too much data in most cases, as this would mostly occur in extreme cases where a smaller of larger z-score would be needed for a more accurate outlier detection and removal.

4.2.3 Feature extraction

All intervals per hour are saved as its length in data points, L_{dp}. To convert the length to milliseconds, L_{ms}, we apply the following formula.

$$L_{ms} = \frac{L_{dp}}{125} * 1000 \tag{4.2}$$

After converting all intervals from data points to milliseconds we will extract 7 different features from the data, x_1,x_2,...,x_7, which will be the input variables for out model.

Mean heart rate

The mean heart rate M in beats per minute of each hour is calculated as follows

$$M = \frac{60000}{\sum_{i=1}^{n} X_i/n} = x_1 \tag{4.3}$$

where X_i is the ith interval, n the total amount of intervals in the hour and 60000 the amount of milliseconds per minute. In many cases the mean heart rate can give a good estimate of a patient's well being. Each age has an interval of beats per minute deemed healthy when at rest and a big diversion out of this interval can already trigger a healthcare worker to take action. As such the mean heart rate can be a key indicator in many cases.

Kurtosis

Kurtosis, k, is often improperly used to described peakedness [90]. Kurtosis actually says something about the combined weight of the tails of a distribution relative to the rest of a distribution and is calculated as the forth moment of the distribution

$$K = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^4}{ns^4} = x_2 \tag{4.4}$$

where X_i is the ith interval, n the total amount of intervals and s the standard deviation of the intervals. A high peak and thin tails of an RR-interval distribution means that the heart rate changed very little during the hour, and vice versa. This means that there was a low overall HRV. This measure however does not take the ordering of intervals and successive changes into account and can thus be seen as a simplified measure of HRV.

Skewness

The skewness, S, calculates the (a)symmetry of a distribution and is calculated as the third moment of the distribution

$$S = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^3}{ns^3} = x_3 \tag{4.5}$$

where X_i is the ith interval, n the total amount of intervals and s the standard deviation of the intervals. The skewness is the asymmetry over the entire distribution without taking the ordering of intervals into account. Asymmetry over the entire distribution can in cases signal a change in heart rate.

RMSSD

The first measure of HRV we use if the Root Mean Square of Successive Differences (RMSSD) of the intervals.

$$HRV = \sqrt{\frac{\sum_{i=1}^{n-1}(X_i - X_{i-1})^2}{n-1}} = x_4 \tag{4.6}$$

where X_i is the ith interval and n the total amount of intervals in the hour. The RMSSD is the most popular HRV measure and has often shown to be a good indicator of a patient's well being as discussed in section 3.3 and section 3.4.

NN50

The second HRV measure we use is the amount of successive RR-intervals that differ by more than 50ms. The distance between R-peaks is also known as an NN-interval and although in detail they might be different, where the NN-interval would exclude certain abnormal peaks, in practise the terms are interchangeable.

$$NN50 = \sum_{i=1}^{n-1}\{\|X_{i+1} - X_i\| > 50\} = x_5 \tag{4.7}$$

where X_i is the ith interval and n the total amount of intervals in the hour.

Guzik

The Guzik Index (GI) calculates as ratio of the difference in distance to the diagonal of a Poincare plot for points above the diagonal over the distance of all points to the diagonal.

$$GI = \frac{\sum_{i=1}^{o} d(P_i)}{\sum_{i=1}^{n} d(P_i)} = x_6 \tag{4.8}$$

where

$$d(P_i) = \frac{\|X_{i+1} - X_i\|}{\sqrt{2}} \tag{4.9}$$

is the distance from point P_i on the Poincare plot to the diagonal, n the total amount of points and o the number of points above the diagonal.

Porta

Porta's index (PI) calculates asymmetry as the ratio of points under the Poincare plot's diagonal over all points that are not on the diagonal. This differs from Guzik's index by only taking the placement of the points with respect to the diagonal into account and not its distance to the diagonal.

$$PI = \frac{\sum_{i=1}^{n-1}\{\|X_{i+1} - X_i\| < 0\}}{\sum_{i=1}^{n-1}\{\|X_{i+1} - X_i\| \neq 0\}} = x_7 \tag{4.10}$$

Where X_i is the ith interval and n the total amount of intervals in the hour.

IDR

The IDR calculates the difference between the first and the ninth deciles, denoted as D_1 and D_9

$$IDR = D_9 - D_1 = x_8 \tag{4.11}$$

Where the minimum and maximum interval length can be heavily influenced by the outlier removal, the IDR is less affected. This makes it a good way to look at the range of a distribution.

4.3 Data labeling

The *PATIENTS* table provides patient specific data including the date of birth and date of death, both for in hospital or thereafter if the death occurred in the period the data collection was still ongoing. Any patients who died after being discharged will be labeled as discharged regardless of the time between discharge and death. Since we are only looking at the last ICU stays of patients in case of multiple stays, we will label any hour from patients who died in the hospital as class label 1, and all other patients label 0. This leaves us with an unbalanced dataset of 64252 hours where 11533 (18%) hours are labeled as 1 and the remaining 52719 (82%) as 0.

4.4 Diagnosis

By taking the ECG of all patients the predictive model will be generalised over multiple diagnoses and ages. Because ECG data records the heart we decided to split patients with any heart related diagnoses from the rest, as they would most likely have the most diverging ECG data from the population average. The MIMIC-III database uses International Classification of Diseases, Ninth Revision, (ICD9) codes to document a patient's diagnosis. The *D_ICD_DIAGNOSES* table consists of 14567 ICD9 codes, including a short and a long title. To find all heart related diagnoses the codes are filtered on keywords in either the short or the long title. The keywords used are the following

- heart
- atrial
- cardiovascular
- myocardial
- coronary
- cardiac
- ventricular

Using these keywords a total of 191 ICD9 codes were identified as diagnoses related to the heart and can be seen in section A.2. When filtering by these codes we found that 2435 patients had heart related diagnoses and 1918 non heart related. This split will be made when creating training and testing sets for our patient specific models, but can be viewed as one group when calculating the ECG features.

4.5 Model creation

An important aspect to consider when choosing a model is interpretability. The reason is twofold; first off, it is important to retrieve what features are deemed important for possible feature selection and future work. This can give insight to researchers on what is important and where they can possibly find improvements. The second reason is for real world usage. When users, in our case healthcare workers, do not know how a model works they are less likely to trust it and therefor also less likely use and apply it. By being able to retrieve what variables were important and why when making a new classification, this information can be translated and shown to the model user. This can both increase trust as well as provide further support when using the prediction to determine what patient's might need extra help. Therefore a Random Forest Classifier (RFC) will be used to predict if a patient will deteriorate or recover in the next 24 hours. A RFC is a white box model where we can directly see how a classification is made and what features were important for each single instance. The ECG features derived from the vital sign data will be used as input and the output will be either class 0 or class 1 as labeled based on in hospital deaths.

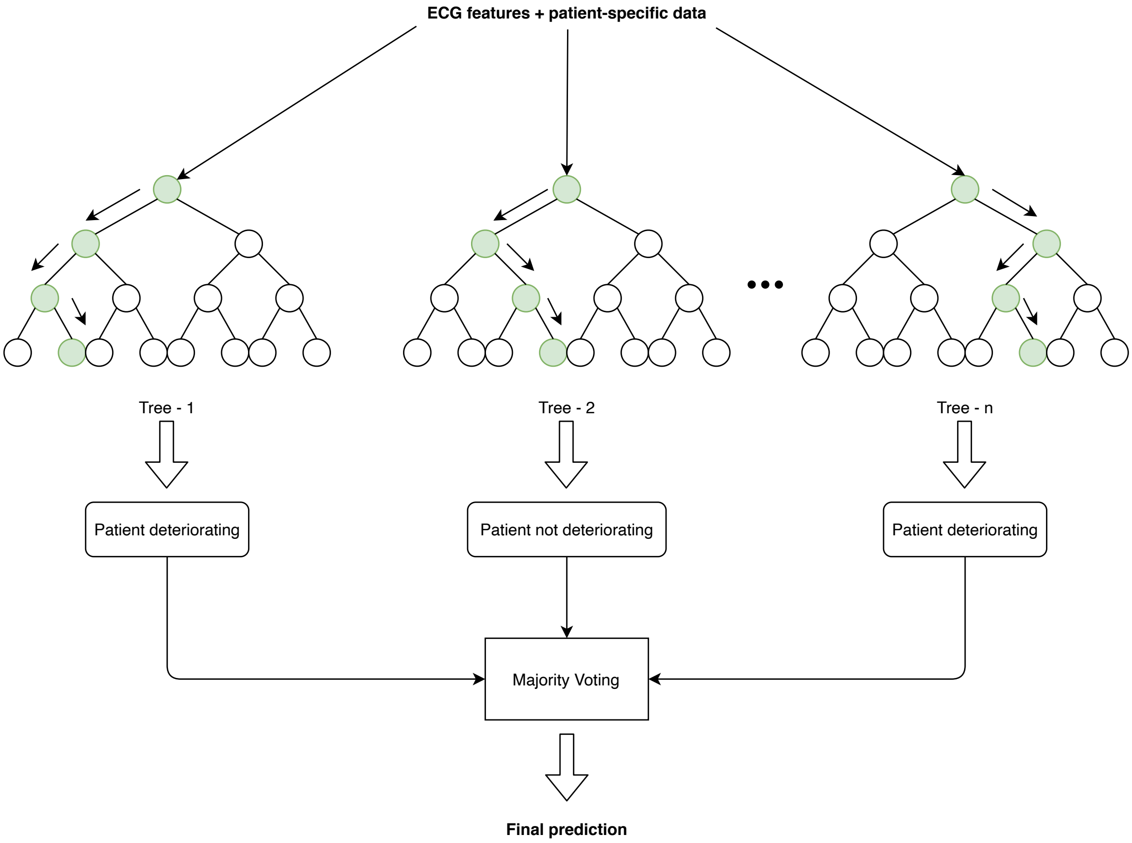

Figure 4.17: Example of a random forest classifier.

A random forest consists of a number of decision trees. Decision trees are easy to interpret and use, but are hard to generalize as they tend to quickly overfit on training data and have trouble correctly classifying new instances. A random forest uses a large collection of decision trees to counter this problem. By using bootstrapping, the random forest will randomly select samples to create a bootstrapped dataset (i.e. bagging) and use a specified number of variables from this dataset to train a tree. This process is repeated for all trees in the random forest. Once completed a new instance can be used as input for the RFC and each tree will classify it according to the variables they are trained on. A majority vote is then used to get to a final classification. An example of this is shown in Figure 4.17.

4.6 Performance evaluation

Different model parameters will be tested to optimise the performance of the model. The performance of a model is often denoted by the accuracy on testing data, the precision, which represents the fraction of correctly classified cases over all classified cases per class. Or by the area under the receiver operator characteristics curve (AUC), which plots the true positive rate (TPR) against the false positive rate (FPR) from 0.0 to 1.0. The model also calculates an out of the bag (OOB) score, which is retrieved by evaluating a tree on the data samples that were left out when bootstrapping.

Since we are predicting patient deterioration and misclassifying a patient who is deteriorating can have fatal consequences we will be looking at more metrics than just the accuracy of the model. One important metric is the recall, also known as sensitivity, which is the fraction of correctly classified cases among all cases for a specific class. A high accuracy with a low recall score means that a lot of patients who were deteriorating were misclassfied even though those who were classified as deteriorating often in fact were. The reverse is the case if most of the people who were deterioration were correctly classified, but also patients who weren't deteriorating were classified as if they were, resulting in a high recall but low accuracy.

In terms of the health of a patient it is important to correctly classify any patient that is deteriorating, thus a high recall is very important. In terms of resource allocation however it is important that not too many patients are classified as deteriorating when they in fact weren't. Both the accuracy as well as the recall are thus important.

The fact that the classes are heavily imbalanced shows that overall only a small amount (in our case 20%) is actually deteriorating. As a result we can assume that a somewhat lower accuracy of class 1 will not increase the amount of false positives by too much and thus a higher recall is deemed more important. To confirm this we will look at the ROC and precision-recall curves in all cases and will aim to optimize this areas under the curves while focusing on achieving a high recall score. We will also look at the effect of under and over sampling of the data using a random under sampler, Near Miss, random over sampler and SMOTE, on the performance of the model as well as introducing class weights.

After creating and evaluating the model we will try to further improve by adding some patient specific variables like the age and/or gender of a patient. This information can make the model input less generic as now patients of different ages and genders can be classified accordingly. This will hopefully enable the model to predict more patient specific cases while not requiring a large amount of patient data, as this data and information can often be missing in low resource settings.

We will split the data in patients with a heart related diagnoses and patients without and will train and evaluate two separate models as well as the effect of introducing a variable that reflects this split in diagnosis. The outcomes will be compared with the main model that takes all data as input to determine if heart related diagnoses significantly differ from other cases by comparing model performance.

As a final evaluation we will look at different times before the end of stay and see if there is a trend in model performance over time, as this can give extra information about the patient's well being and if he or she is recovering or not.

4.7 Feature importance

Single decision trees can often be graphically represented if they aren't too big (e.g. depth and number of variables). A random forest however often contains hundreds of trees and can therefor not be easily visually interpreted by a human anymore. But as the structure can still be read we can see what nodes/decisions are used to get to a prediction and thus calculate the importance of a feature. This can be done on the model level, where each tree is analyzed and the information gain at each node for each feature can be calculated. Another way is by following a single instance as it goes through the model during classification and again looking at the information gain. This is called the observation level feature importance and can be applied in real time when using the model in an application. We will consider both types of feature importance in this thesis.

Chapter 5

Experiments

5.1 Summary

In the following sections we will be fitting, evaluating and optimizing random forest classifiers. For the evaluation of the model we will keep track of the AUC, precision, recall and OOB scores. Apart from the precision and recall scores by themselves we also look at the area under the precision-recall curve, which shows the trade-off between precision and recall, where a high area means both a high precision as well as a high recall.

To start we will train a model using default parameter values and will optimise the most important features using a grid search. Once we found the best performing parameter values we will try under and oversampling techniques and add class weights to deal with the data imbalance and further improve the model performance.

We will then try to obtain an even higher model performance by adding patient specific variables to our input dataset. This addition will be tested on two previously best performing models based on different scores. For patients with heart related diagnoses we make a separate subset as well as a new input variable and we evaluate and compare the performance of using separate subsets and models versus using the diagnoses class as an input variable.

After these experiments we continue with the two best performing models and test them in series by using the positively classified subset from one model as input for the next and evaluate the effect on the true and false positive rates.

Another metric that will be obtained is the performance and classification probability over time, ranging from the first to the last hour of the patient's ICU stay.

The two best performing models will then be taken under the loop and feature importance on both model as well as observation levels will be extracted and evaluated.

Finally we will look at the effect of using different threshold levels on the classification performance of one of our models with the aim to get a more balanced output.

5.2 Base model

5.2.1 Default parameter values

To start we trained a Random Forest Classifies with its default settings as given by the Sklearn library [67] on all the data. The main parameters of these settings are shown in Table 5.1.

Parameter	Value	Description
max_depth	None	The max depth of a tree
n_estimators	100	The number of trees
min_sample_split	2	The minimum number of samples required to split an internal node
min_sample_leaf	1	The minimum number of samples required to be at a leaf node

Table 5.1: Most important model parameters and values

Note that 'None' is the default parameter for the max depth of a tree and means that nodes are expanded until all leaves are pure or until all leaves contain less than min_samples_split samples. The data is split in 80% training (51401) and 20% testing (12851) and the class ratio is kept the same over both subsets as seen in Table 5.2.

Subset	Class count	Percentage
Training	Class 0: 42175 Class 1: 9226	Class 0: 82% Class 1: 18%
Testing	Class 0: 10544 Class 1: 2307	Class 0: 82% Class 1: 18%

Table 5.2: Training and testing subsets

The model was fitted on the training data and afterwards evaluated on testing data. On top of this the OOB score was also tracked. The results are shown in Table 5.3.

	Precision	Recall
Class 0	0.85	0.99
Class 1	0.76	0.18

Table 5.3: Base model performance using default parameter values

The OOB score was 0.84, which is in line with the precision of the model and reached an AUC of 0.77, as seen in Figure 5.1. When looking at the recall however we notice that class 1 scores very low. This means that a lot of patients who were deteriorating have been missed. The reason why the precision and AUC are still fairly high stems from the imbalance in the data and the model overfitting on class 0.

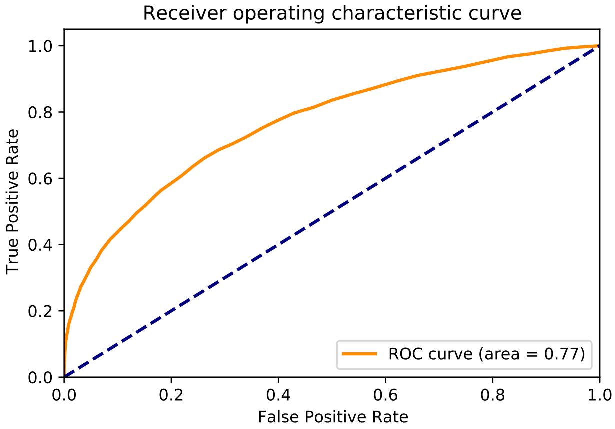

Figure 5.1: ROC curve of the base model using default parameter values

This was also resulted in a low area under the precision-recall curve of class 0 of only 0.494, whereas class 1 reached an area of 0.927, as seen in Figure 5.2.

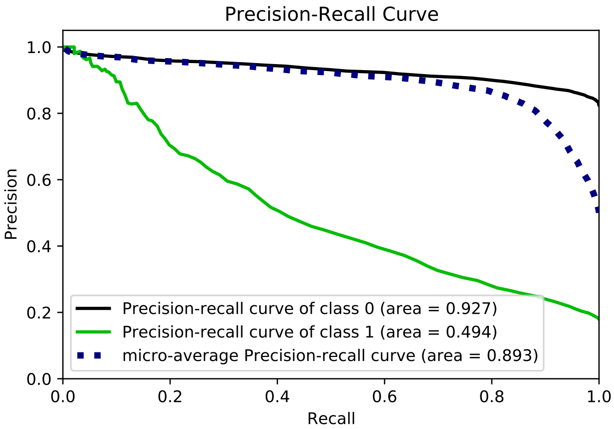

Figure 5.2: Precision-recall curve of the base model using default parameter values

5.2.2 Optimization

To increase model performance we ran a grid search on the main parameters of the RFC. The grid search tries all combinations of parameter values given. The value ranges are shown in Table 5.4.

Parameter	Min value	Max value	Step size
max_depth	2	50	8
n_estimators	500	2000	150
min_sample_size	2	5	1
min_samples_leaf	1	4	1

Table 5.4: RFC parameter value grid

The grid search uses 3-fold cross validation and fits the model on the same dataset. The best parameters were a max_depth of 50, a n_estimators of 650, min_sample_size of 2 and min_sample_leaf of 1. This combination of parameters resulted in a precision of 0.78, an AUC of 0.78 and an OOB score of 0.84. What stands out is that most of the tweaked parameters are not far from its default value except for the max_depth, which was best at its max. Looking at the scores in Table 5.5 however we can see that this depth did not increase the performance significantly, while greatly adding to the complexity of the model.

	Precision	Recall
Class 0	0.85	0.99
Class 1	0.78	0.18

Table 5.5: Base model performance using grid search's best parameter values

To see the effect of the max_depth of the model we ran the optimal parameters with an increasing max_depth from 1 till 50 and plot the precision and recall in Figure 5.3 to see at what depth we already find a similar score with the aim of using a smaller depth as this would reduce the model complexity.

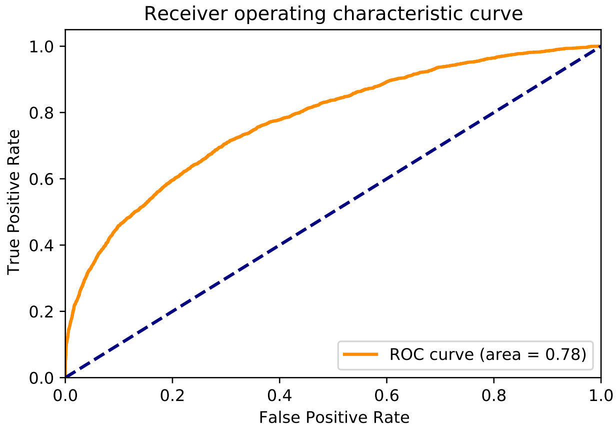

Figure 5.4: ROC curve of the base model using optimized parameter values

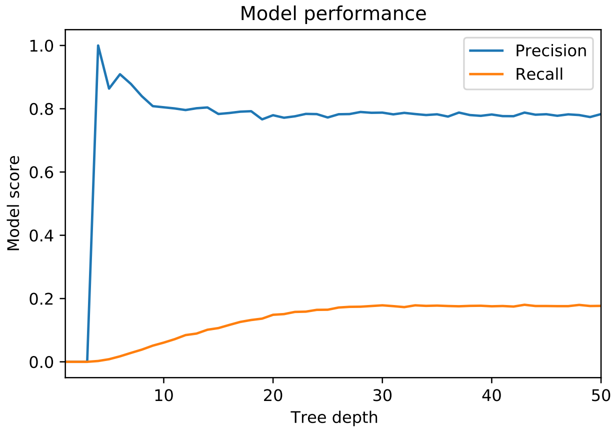

Figure 5.3: Precision and recall at different depths

We can see that the recall reaches its limit at a depth of 30 and performs as good as with a depth of 50. The final parameter changes after optimisation are thus a max_depth of 30 and using 650 estimators.

Using these parameter values we fitted and evaluated the model and compared the AUC as well as the precision-recall curve with the one from the base model. The AUC slightly increased from 0.77 to 0.78, while on the precision-recall curves the class 1 curve area increased from 0.494 to 0.514 and that of class 0 from 0.927 to 0.931, as seen in Figure 5.4 and Figure 5.5. The OOB score remained the same at 0.84. Overall the model performed slightly better on these optimised parameters.

5.2.3 Dealing with unbalanced data

Although the model's precision is fairly high, the recall of class 1 is too low for any real world usage. The reason for this is the imbalance of the data, making the model favour classifying instances as class 0 over class 1. To counter this we will use both under as well as over sampling techniques. We are using both because with undersampling only original data will be used in the final training set, where as with oversampling new instances will be created the were either not in the original dataset or are duplicates of original instances and therefor not adding any new data

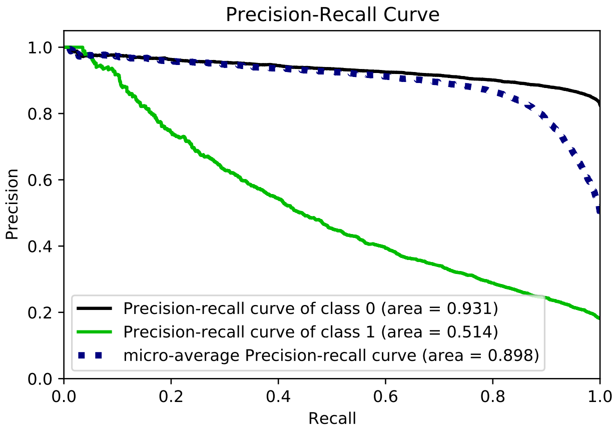

Figure 5.5: The precision-recall curve of the base model using optimized parameter values

to learn from. Class weights will also be used to try and counter the class imbalance and model bias.

Under sampling

One way of countering class imbalance is by under sampling the majority class. For this the Random Under Sampler (RUS) and Near Miss from IMBLEARN [48] are used.

The RUS method randomly selects and removes instances from the majority class until the two classes are of the same size, i.e. 9226. The model is trained on the new balanced subset but still tested and evaluated on the unbalanced testing set. The accuracy as seen in Table 5.6 is significantly different, most noticeably the precision of class 1 decreased while increasing the recall. The decrease in precision is also noticeable in the OOB score which dropped to 0.68.

	Precision	Recall
Class 0	0.91	0.67
Class 1	0.32	0.70

Table 5.6: Model performance on undersampled data using RUS

In our case this can be a more desirable outcome as only 30% of the deteriorating patients are missed. One third of the recovering patients were classified as deteriorating, meaning an extra 3480 patients would have gotten extra care when not needed if applied in a real world setting.

The difference in performance did not affect the AUC much as it only slightly decreased from 0.78 to 0.76. The steep drop in precision did however result in a worse precision-recall curve as the area was only 0.457.

The second method, Near Miss, selects instances to remove by first calculating the distance between instances of both classes and selecting the ones of the majority class closest to the minority class to be removed first, eliminating the instances that were a ”near miss”. However hen applying this method to our data, the model heavily leans towards class 1, reaching a high recall but low accuracy as seen in Table 5.7, and an OOB score of 0.86. Although this means that most deteriorating patients are correctly classified it does also misclassify most instances of class 0, making the model not useful for real world usage. One possible reason for this is that class 0 instances might be more spread out. Class 1 instances in the middle between such class 0 instances would be filtered out while being relatively further separated from class 0 compared to instances that are close to class 0 but not in between. This and other specific class distributions can cause near miss to not perform optimal.

	Precision	Recall
Class 0	0.87	0.27
Class 1	0.20	0.82

Table 5.7: Model performance on undersampled data using Near Miss

As a result the AUC dropped to 0.54 and the precision-recall area dropped all the way down to 0.19 for class 1, with 0.376 for the micro average curve area.

Over sampling

Another technique of balancing the data is oversampling the minority class. We trained two separate models using two oversampling techniques, namely Random Over Sampling (ROS) and synthetic minority oversampling technique (SMOTE).

ROS creates new samples of the minority class that are close to the existing instances or in some cases duplicates. It does this until both classes are of equal size, i.e. 42175. After evaluating on the testing data the scores are retrieved as seen in Table 5.8. The OOB score increased to 0.98.

	Precision	Recall
Class 0	0.86	0.97
Class 1	0.66	0.26

Table 5.8: Model performance on oversampled data using ROS

The difference compared to the base model is mostly noticeable in class 1 as the precision went down and the recall went up while not seeing a significant difference in class 0 and having a AUC of 0.78. Compared to the optimised base model the precision-recall curves are almost identical, with a 0.515 area under the curve for class 1 and 0.932 for class 0.

SMOTE is a commonly used oversampling technique which looks at the k-nearest neighbors of an instance in the minority class and creating new samples by linear interpolation in this space. The default value for k is 5 and to see the effect of different k values we ran the over sampler with values ranging from 2 to 25. The corresponding precision and recall are plotted in Figure 5.6 and it is clearly visible that the k-value has little to no impact on the model score.

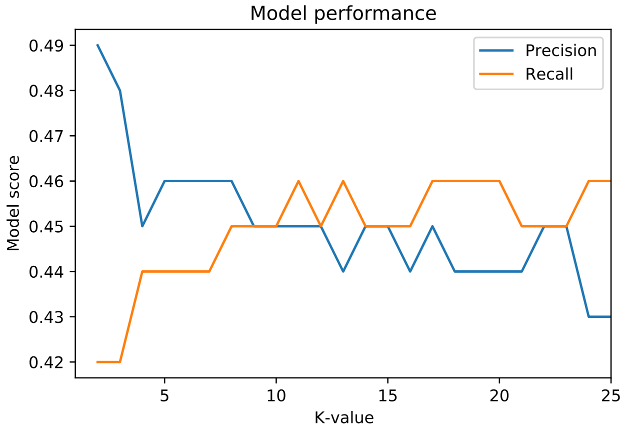

Figure 5.6: Model precision and recall after oversampling using SMOTE with varying k-values

The best score was achieved using k=8 and the model performance is shown in Table 5.9 and an OOB score of 0.88.

	Precision	Recall
Class 0	0.88	0.88
Class 1	0.46	0.45

Table 5.9: Model performance on oversampled data using SMOTE with k=8

Using SMOTE with k=8 resulted in an AUC of 0.76, slightly lower compared to ROS. Also the precision-recall was slightly worse when using SMOTE, with the class 1 area under the curve dropping to 0.473 while class 0 stayed almost the same at 0.926.

Class weights

A way of accounting for class imbalance without changing the dataset in any way is by introducing class weights. To create an equal weight we will weigh our classes 1:4, with class 1 weighing 4 times as much as class 0. The RFC takes these weights into account when fitting to the data and will try to optimize the precision. A higher precision is desired but as discussed in previous sections the recall is of bigger importance and as seen in the results in Table 5.10 the recall has decreased even further when comparing to the base model and a slightly lower OOB score of 0.84 was scored.

	Precision	Recall
Class 0	0.84	0.99
Class 1	0.79	0.15

Table 5.10: Model performance using class weights

Adding class weights does not alter the actual amount of instances per class and as a result the increase in class 0 classifications did not result in a change in either the AUC or the precision-recall curves, with both reaching the same areas under the curve as without the class weights.

5.3 Patient specific model

The imbalance in the data can reasonably be countered by undersampling resulting in a more favorable accuracy for a real world setting. It is however still lower than desired as almost a third of the deteriorating patients is still missed and a fairly large amount of recovering patients are wrongly classified. This model did however decrease the area under the precision-recall curve compared to our optimised base model and as such both models will be used for further optimisation. The main reason for these model outcomes is the population the data is gathered from. By generalising over all patients of all ages 18-88 and all diagnoses the model can be theoretically applied in settings were patient information might be missing. Adding a few patient specific constant variables to the model could however increase the performance significantly as on top of the differences in ECG features the instances could be further distinguished and split by the RFC. To not increase the requirements for the model too much we only added easy to gain information, namely a patients age and gender.

5.3.1 Adding gender

Original data

We first fit and evaluate a model which includes the patients gender without augmenting the data. We use the same model parameters to have a clear comparison between models. The results and comparison to our optimised models without gender are shown in Table 5.11 where we see a small increase in class 1 precision. Also the OOB score slightly increased to 0.85.

	Precision	Recall	Precision diff	Recall diff
Class 0	0.85	0.99	0.00	0.00
Class 1	0.82	0.18	+0.04	0.00

Table 5.11: Model performance and comparison when including patient's gender

As the results were slightly better than the original model the AUC increased a little to 0.79. Also the precision-recall area for class 1 and 0 increased to 0.539 and 0.937 respectively.

Undersampled data

The same evaluation is performed on a model fitted with balanced data including gender using RUS. As seen in Table 5.12 the model performance slightly increased compared to the model train using RUS in the previous section, with the OOB score also slightly increasing to 0.69.

	Precision	Recall	Precision diff	Recall diff
Class 0	0.91	0.68	0.00	+0.01
Class 1	0.33	0.71	+0.01	+0.01

Table 5.12: Model performance and comparison when including patient's gender and balancing the classes using RUS

As the results weren't very different from the original model the AUC did not change, reaching the same 0.77. The slight increase in precision and recall for class 1 are however visible in the area under the precision-recall curve, increasing from 0.457 to 0.48.

5.3.2 Adding age

Next we look at the effect of adding the patient's age as an input. The results are again compared to our optimised models without age using both the original data as well as undersampled data.

Original data

The model performance using the data including the patient's age is shown in Table 5.13. Including the patient's age has a significant effect on model performance, increasing the class 1 precision and recall greatly as well as the OOB score which increased to 0.87.

	Precision	Recall	Precision diff	Recall diff
Class 0	0.87	0.99	+0.02	0.00
Class 1	0.90	0.32	+0.12	+0.14

Table 5.13: Model performance and comparison when including patient's age

This increase is clearly visible in the AUC, increasing the score from 0.78 to 0.88, while also increasing the area of class 1 precision-recall curve from 0.514 to 0.703 and class 0 to 0.966.

Undersampled data

The same evaluation is again performed on a model fitted with balanced data including age using RUS and compared with the optimised model on data without the patient's age and using RUS. As seen in Table 5.14 the model performance increase all over the board and seeing an increase of OOB score to 0.76.

	Precision	Recall	Precision diff	Recall diff
Class 0	0.94	0.74	+0.03	+0.07
Class 1	0.40	0.77	+0.08	+0.07

Table 5.14: Model performance and comparison when including patient's age and balancing the classes using RUS

The model also got an AUC of 0.85, significantly higher than any previous model at 0.76. The increase didn't translate to the precision-recall curve where the area for class 1 decreased from 0.688 to 0.621 and for class 0 slightly from 0.967 to 0.958.

5.3.3 Adding gender and age

Both gender and age showed improvements in model performance. We tested if adding both variables as input to the model would show significant improvements over adding either one.

Original data

Perhaps unsurprisingly adding both gender and age as input increased the performance of the model even further, as seen in Table 5.15. This inclusion also improved compared to the previous best model using only age as the precision and recall of class 1 increased by another 0.02 and 0.01 respectively. The OOB score slighly increased as well 0.88.

	Precision	Recall	Precision diff	Recall diff
Class 0	0.87	0.99	+0.02	0.00
Class 1	0.92	0.33	+0.14	+0.15

Table 5.15: Model performance and comparison when including patient's age and gender

This increase also improved the AUC to 0.90, reaching an area under the precision-recall curve of 0.738 for class 1 and 0.972 for class 0.

Undersampled data

This increase in performance is also noticeable in the model after using RUS to balance the data when both variables are added, as shown in Table 5.16. This means that only 21% of deteriorating patients were missed while only 24%, or 2530, patients were misclassified as deteriorating while they were not. The model scored an OOB score of 0.77 using this method.

	Precision	Recall	Precision diff	Recall diff
Class 0	0.94	0.76	+0.03	+0.09
Class 1	0.42	0.79	+0.10	+0.09

Table 5.16: Model performance and comparison when including patient's age and gender and balancing the classes using RUS

The AUC also went up to 0.86 while the precision-recall curve improved and reached an area under the precision-recall curve of 0.645 for class 1 and 0.962 for class 0.

5.3.4 Heart related versus other diagnoses

When using ECG data as input it means the heart is the key to making the prediction. Naturally speaking if someone has a heart condition this can result in different ECG data and thus different feature values. To further optimise the performance, the data is split into two subset; one of patients who had a heart related diagnosis and one without. Two models were fitted for each subsets and evaluated on the models having age and gender as an input as described in the precious section.

Heart related diagnosis

After filtering, the subset for patients with a heart related diagnosis contains a total of 35770 hours of data, of which 23288 class 0 and 5328 class 1 in the training subset.

Original data

First we fit and evaluate the model without altering the data. The performance and comparison with the previous best performing model is shown in Table 5.17, got an OOB score of 0.89 and improved mostly for class 1.

	Precision	Recall	Precision diff	Recall diff
Class 0	0.88	0.99	+0.01	0.00
Class 1	0.94	0.39	+0.02	+0.06

Table 5.17: Model performance and comparison when fitting and evaluating on patients with heart related diagnoses

The AUC improved slightly from 0.90 to 0.91. Also the precision-recall curve improved to 0.782 for the class 1 area and stayed roughly the same at 0.973 for class 0.

Undersampled data

We also undersampled class 0 and fit the model as in the previous sections. The result as shown in Table 5.18 shows again an increase in performance, compared to the previously best performing model after adding age and gender and using RUS, and achieving an OOB score of 0.80.

	Precision	Recall	Precision diff	Recall diff
Class 0	0.94	0.79	0.00	+0.03
Class 1	0.47	0.80	+0.05	+0.01

Table 5.18: Model performance and comparison when fitting and evaluating on patients with heart related diagnoses and balancing the classes using RUS

The increase in performance also resulted in an increase in AUC, reaching 0.88, and as a result of the increased performance also the precision-recall curve improved to 0.708 for class 1 and 0.964 for class 0.

Non heart related diagnosis

The data that is left belongs to all patients that did not have any heart related diagnosis. Although this is still a mixture of many different diseases and cases, the difference in ECG data and features can be expected to be less extreme compared to those with heart related diagnoses. In total this subset consists of 28482 hours of data, of which 18887 hours of class 0 and 3898 hours of class 1 in the training set.

Original data

The model performance and comparison are shown in Table 5.19. The most notable change is the increase in recall for class one. The OOB score for this model was 0.89.

	Precision	Recall	Precision diff	Recall diff
Class 0	0.89	0.99	+0.02	0.00
Class 1	0.92	0.43	0.00	+0.10

Table 5.19: Model performance and comparison when fitting and evaluating on patients without heart related diagnoses

Although some scores increased significantly it did not result in a much higher AUC, which was up 0.01 from that of the model considering all patients at 0.91. As a result of the increased performance the precision-recall curve did improve significantly as the area for class 1 increased by 0.045 to 0.783 and the class 0 area with 0.007 to 0.979.

Undersampled data

As a final experiment to increase the model performance and real world applicability data was undersampled using RUS and a model was fitted and evaluated agains the previously best performing model. The results on this subset when using RUS, as shown in Table 5.20, show again an increase in performance as well as in OOB score, which was 0.80 for this model.

	Precision	Recall	Precision diff	Recall diff
Class 0	0.95	0.80	+0.01	+0.04
Class 1	0.45	0.80	+0.03	+0.01

Table 5.20: Model performance and comparison when fitting and evaluating on patients without heart related diagnoses and balancing the classes using RUS

This difference is also noticed in the AUC, which increased by 0.02 to 0.88. Finally the precision-recall curve also improved to an area for class 1 of 0.688 and a class 0 area of 0.970.

Diagnosis as input

In the previous section we have seen that splitting the data by diagnoses improves the model performance. This does however double the amount of models needed to account for each possible combination of wanted output and patients diagnosis. A way to combine this is by using an input variable to denote if a patient has a heart related diagnosis or not. The variable *HEART_RELATED* will be used for this and is 1 for patients with a heart related diagnosis and 0 otherwise. A model is trained and evaluated on both the original data (Model 1) including this variable as well as balanced training data using RUS (Model 2) and compared to the averages of the two models for both diagnosis groups in the previous section.

As seen in Table 5.21 The performance when separating by diagnosis using a variable is nearly identical to the average of using two separate models when using the original dataset except for the decrease in class 1 recall, which was already very low. Furthermore the model got an OOB score of 0.88, a slight decrease.

	Precision	Recall	Precision diff	Recall diff
Class 0	0.88	1.00	-0.01	+0.01
Class 1	0.94	0.36	+0.01	-0.05

Table 5.21: Model performance and comparison when fitting and evaluating using the heart related diagnosis as an input variable versus using two separate models when using the original dataset

The decrease in model performance was barely noticeable in the precision-recall curve and not at all in the ROC, which actually increased to an AUC of 0.92.

The overall performance thus slightly decreased, but with the aim of not misclassifying recovering patients as to limit resources the class 0 recall and class 1 precision are desired to be high, which the model did achieve. With the aim of not missing any deteriorating patients we want to achieve an as high as possible class 1 recall, which thus far has been achieved when training the model using the RUS, we performed the same model fitting using balanced data and compared it with the average of the previous models that were trained using split data. As shown in Table 5.22 this model also performed slightly worse compared to the average of the previously evaluated models, but not on the most important score, being the class 1 recall, which increased to 0.81. The model received an OOB score of 0.78, slightly lower than the 0.80 from the previous models.

	Precision	Recall	Precision diff	Recall diff
Class 0	0.95	0.78	0.00	-0.02
Class 1	0.44	0.81	-0.02	+0.01

Table 5.22: Model performance and comparison when fitting and evaluating using the heart related diagnosis as an input variable versus using two separate models when using RUS to balance the dataset

The decrease in performance resulted in a the same AUC of 0.88, and only a slight decrease in the area under the precision-recall curve to 0.677 for class 1 and 0.966 for class 0.

5.4 Models in series

As seen in the previous section there is a cost when using the diagnosis as input instead of using separate models. The benefit however is a smaller need for resources, as a single model has a size of up to 500MB and a medical device might have limited capacity. Although exact requirements for this are not known as of writing, the first estimate as given are to stay below 1GB in total size. Given how the decrease in performance was mostly noticed in the less important scores for each model, we will assume that the benefits of only using two models outweigh the costs.

Combining these tho model can result in an even higher score when first classifying all deteriorating patients on Model 2, with high class 1 recall, and further evaluation the outcome by feeding this sub population into Model 1, with a high class 1 precision.

To test this we use the entire testing dataset containing of 12851 hours of data, of which 10544 belong to class 0 and 2307 belong to class 1. As seen in Table 5.23, there were 4211 hours classified as deteriorating, of which 1863 were true positives and 2348 false positive.

	Class 0	Class 1	
Class 0	8196	2348	10544
Class 1	444	1863	2307
	8640	4211	12851

Table 5.23: Confusion matrix after running the testing data through the first model

All hours labeled as class 1 are filtered out of the original test set to create the output subset of the first model. This output subset is then used as input for the second model. The model output is shown in Table 5.24.

	Class 0	Class 1	
Class 0	2297	51	2348
Class 1	1027	836	1863
	3324	887	4211

Table 5.24: Confusion matrix after running the model 1 output subset data through the second model

The second model also performed slightly better on class 1 as it did on the entire test set, with the class 1 recall increasing from 0.36 to 0.45 while keeping the precision at 0.94.

As a result out of 2307 deteriorating patients 836, or 36%, were correctly classified. On the other side out of 10544 recovering patients only 51, or 0.5%, ended up being classified as deteriorating, meaning that in a real world setting where the balance in data is comparable with our dataset (80% recovering versus 20% deteriorating), there are hardly any false positives.

5.5 Performance over time

In a real world setting as patient's heart rate and variability can change a lot from hour to hour and it is therefore also important to look at the change over time. To validate if our models show a trend in predictions following the trend in recovery or deterioration we will evaluate the model on data from 1, 12 and 24 hours before the end of stay. In other words, when a patient recovers or deteriorates, we assume the closer they are to release or death, the more clearly it can be seen in the data, resulting in a higher model performance and a higher certainty closer to the end of stay. We will both look at the model performance on these specific hours as well as the change in probability when classifying. For the change is probability we will look at the mean probability over all positive and negative labels, regardless of them being correctly classified or not, i.e. the probability can be < 0.50 if there are a lot of false negative/positives.

After filtering the testing data on hours before end of stay we get subsets as described in Table 5.25.

Hours till end of stay	Class 0	Class 1	Total
1	377	84	461
12	417	116	533
24	419	83	502

Table 5.25: Data amounts for different times before end of stay

5.5.1 Model 1

Figure 5.7 shows the trend in model performance and prediction probability for both classes at times 24, 12 and 1 hour(s) before the end of stay. No clear upwards trend was found when using the first model.

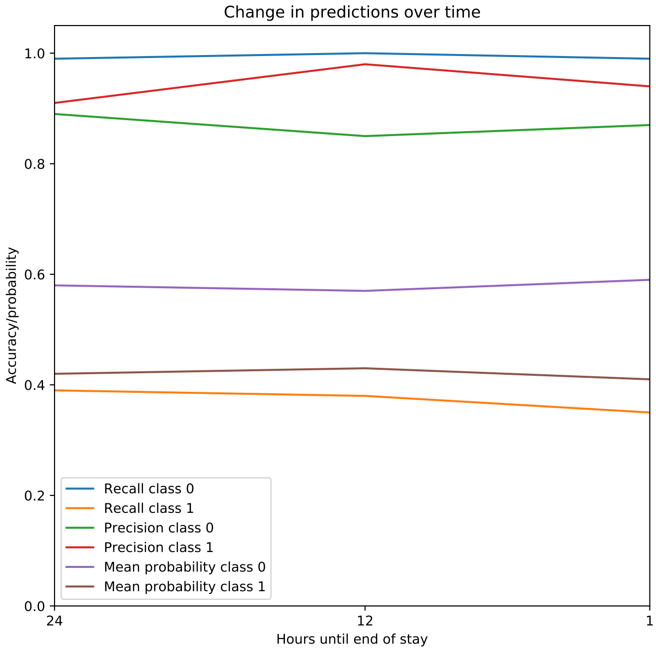

Figure 5.7: Change over time in model 1 performance and prediction probability

This means that the model is not more certain that a patient is deteriorating closer to its death than it is 12 or 24 hours before. A reason for this can be that the chronological order of the hours is not taken into account when fitting the model. Each hour is an instance that is handled individually from the rest and changes per successive hour are not taken into account. However a trend could still be reasonably expected, as a patient should be in a much worse condition close to its death compared to half a day or more before.

5.5.2 Model 2

In model 2 we again see no strong trend, as can be seen in Figure 5.8, although we do see a slight increase in the mean probability when predicting deterioration, which is the key indicator in real world usage, as this tell the healthcare worker how certain the model is that a patient is deteriorating. This outcome is comparable with model 1.

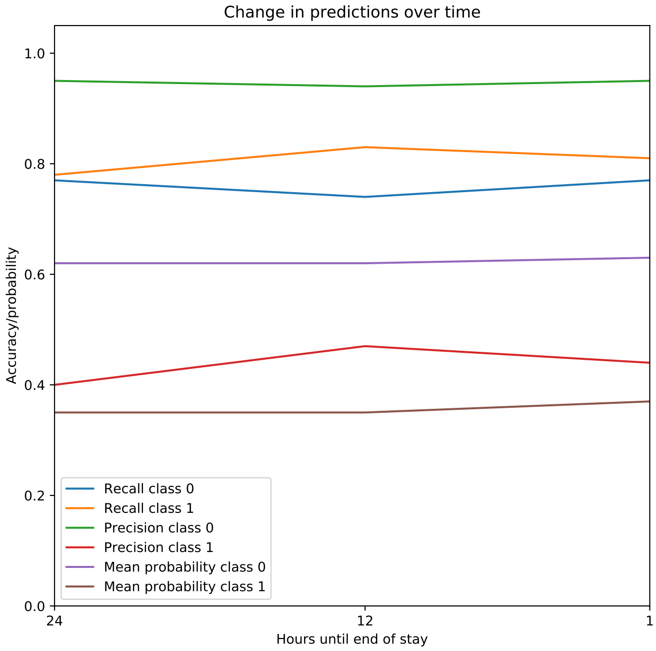

Figure 5.8: Change over time in model 2 performance and prediction probability

5.5.3 Models in series

The key indicator in real world usage is the probability that a patient is deteriorating. As a final check we will look at how the models in series perform on class 1 predictions, looking at the final output. As seen in Figure 5.9 the recall decreases the closer we get to a patients death. Although this means we are missing more patients who are deteriorating, we do see a slight increase in model certainty that these patients are deteriorating.

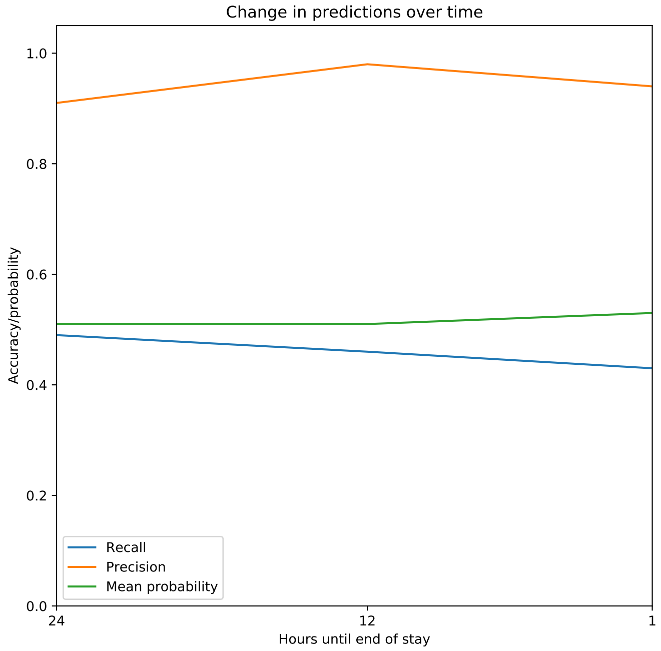

Figure 5.9: Change over time for the models in series performance and prediction probability

5.6 Feature importance

Now that we have two models fitted and evaluated, the next step is to interpret them and retrieve what features are the most important when making a classification.

5.6.1 Model level feature importance

The model level features are directly derived from the model and reflect which features are most important in explaining the output of the model. Note that gender and the heart related diagnosis flag are binary variables while this method can be biased towards continues variable, making these binary variables less important at model level.

We first look at Model 1. As seen in Figure 5.10 the most important features are a patient's age and mean heart rate, while the heart related diagnosis and gender variables are deemed least important. The same can be observed when looking at the feature importance of Model 2 as shown in Figure 5.11.

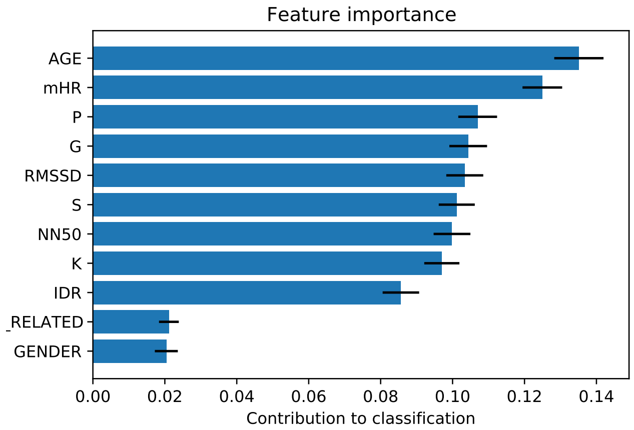

Figure 5.10: Model 1 feature importance at model level

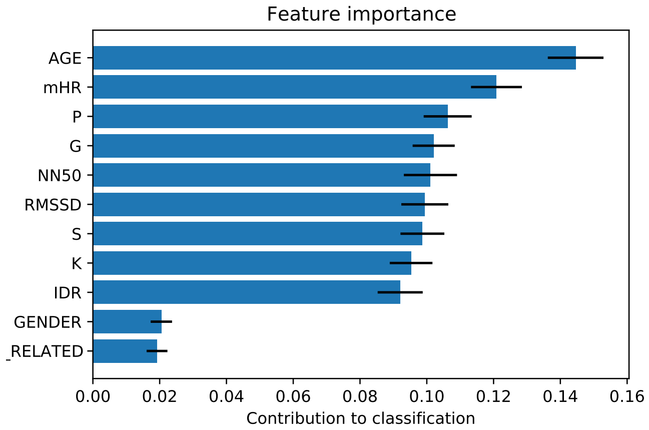

Figure 5.11: Model 2 feature importance at model level

5.6.2 Observation level feature importance

An arguably better way of retrieving feature importance is at the observation level, were we look at the feature contribution when making a classification. For this we use the TREEINTERPRETER python library [1]. The tree-interpreter gives as output the model bias plus the contribution of each feature to the final prediction. This method can be used to e.g. see what features are most important when predicting deterioration, to find out what features contribute most to misclassifications or to what features seem overall less important. On top of this the interpreter can be used in real time to give healthcare workers more insight into what the model bases its prediction on.

Model 1

We first look at model 1 and run the tree-interpreter on the model given the test set. As the model was evaluated on unbalanced testing data it is very biased towards class 0, with a bias of 0.82 versus a bias for class 1 of 0.18. For an instance to be classified as 1 the sum of feature importance has to be ≥ 0.32. To evaluate what features are most important we look at the true positives and retrieve for each instance what feature had the biggest positive impact. An overview of the most important features is shown in Figure 5.12 and shows that the patient's age is by far the most important feature.

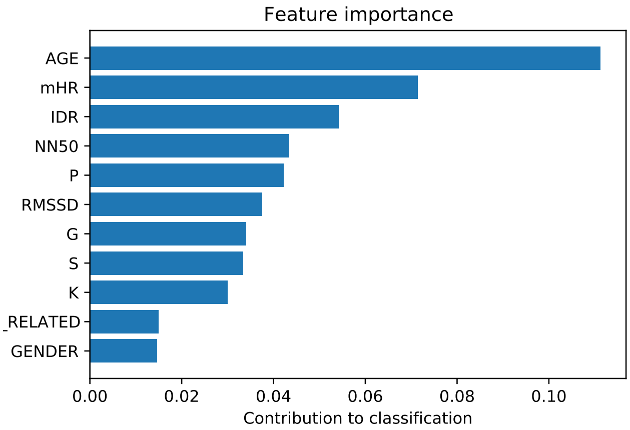

Figure 5.12: Model 1 feature importance at observation level when classifying true positives

Furthermore we look at what variables are most often the leading factor in making a true positive prediction and see that age is the most important feature in over 42% of the cases, as shown in Figure 5.13, followed by the mean heart rate in almost 21% of the cases.

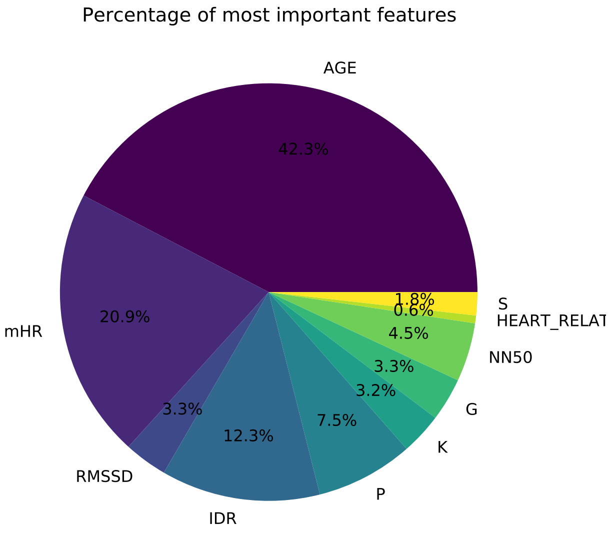

Figure 5.13: Model 1 percentage of feature being leading in making a true positive classification

Another interesting group to analyse are the false negatives, where we can look at what features weigh heaviest in making these misclassifications. In Figure 5.14 we can see that although the features on average all predict a negative importance, meaning that they lean towards the positive class in case of a false negative result, the total feature importance is not enough to outweigh the bias that the model has.

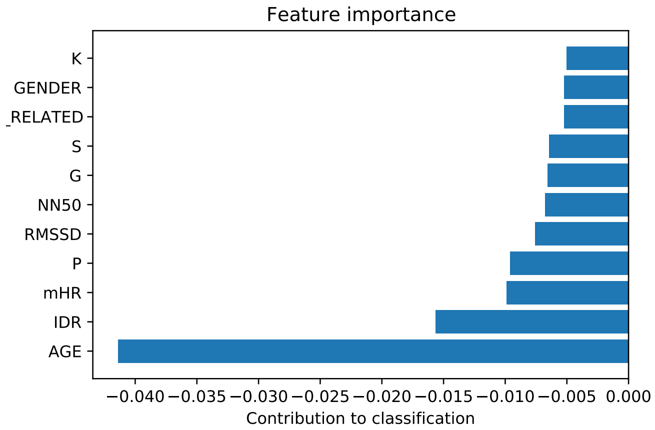

Figure 5.14: Model 1 feature importance at observation level when classifying false negatives

How often it is the main reason for a false negative can be seen in Figure 5.15, where somewhat surprisingly the age of a patient is only a leading factor in 11.3% of the cases while the IDR is in over 18%.

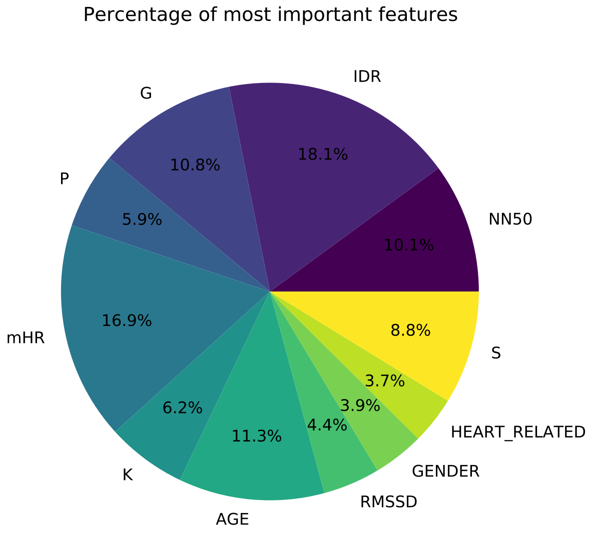

Figure 5.15: Model 1 percentage of feature being leading in making a false negative classification

Model 2

The same test are done for the second model, where we again observe that the age is the most important factor, but follow by the IDR as the second most whereas in the first model the mean heart rate had a higher importance. We do again observe that the binary features for the patient's gender and heart related diagnosis were least important (Figure 5.16).

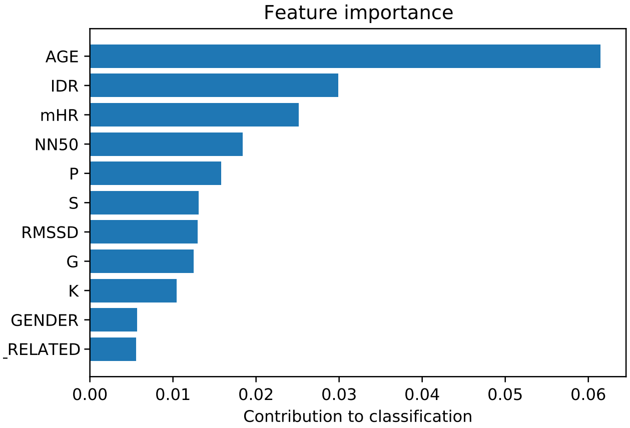

Figure 5.16: Model 2 feature importance at observation level when classifying true positives

This was also reflected in the percentage of classifications each features was deemed most important as shown in Figure 5.17.

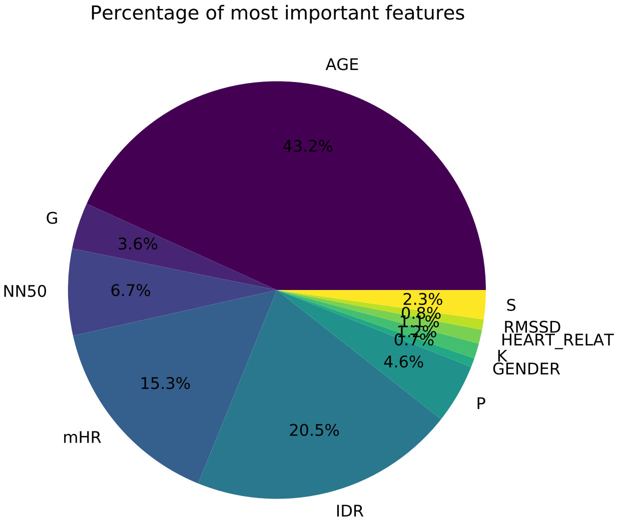

Figure 5.17: Model 2 percentage of feature being leading in making a true positive classification

The same evaluation is done on the false negative prediction. One thing that is clearly visible when looking at Figure 5.18 is that almost all features on average weigh towards the negative class it was classified as. Another interesting finding is that both the gender as well as heart related diagnosis variables do point towards the positive class. This is also the first case where a patient's age wasn't the clearly most important feature, as it scored nearly equal to the mean heart rate and NN50 while the IDR was most important.

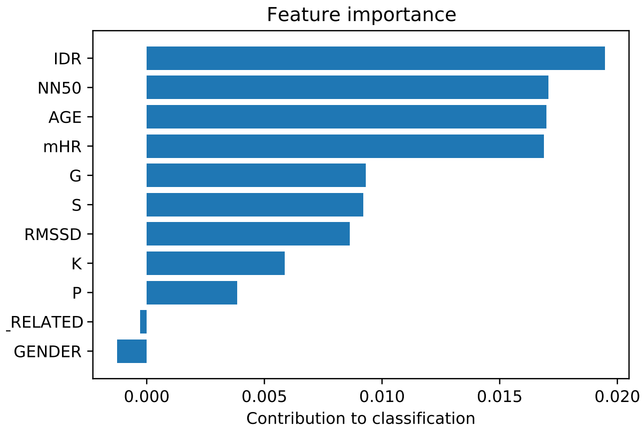

Figure 5.18: Model 2 feature importance at observation level when classifying false negatives

Somewhat surprisingly the average feature importance was not reflected in the percentage of cases where the features were most important in making the classification. In Figure 5.19 we can see that the age was still the most important in most cases, followed by the IDR and mean heart rate. The NN50 was most important in only 10.9% of cases were as on average is was the second most important feature.

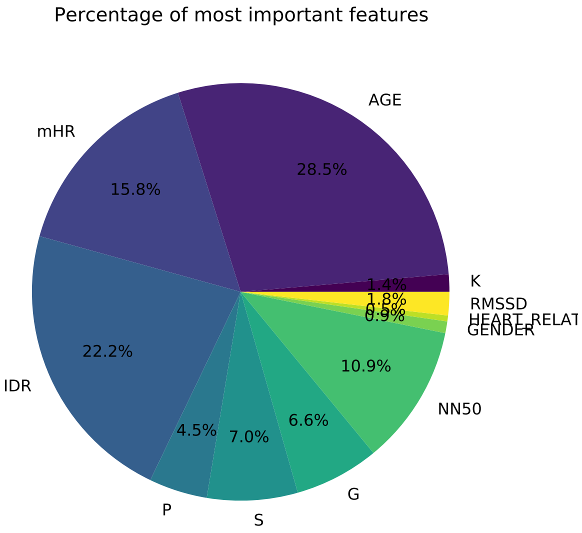

Figure 5.19: Model 2 percentage of feature being leading in making a false negative classification

5.7 Threshold

The default threshold for binary classification is 0.5, meaning that a probability ≥ 0.5 will be classified as 1, and as 0 otherwise. As seen in Figure 5.14, model 1 on average had a negative feature importance on false negative classifications, meaning its features predicted the opposite class, namely positive, but the bias was too big to get a class 1 probability of ≥ 0.5. This bias can be countered by lowering the threshold level needed to classify a prediction as true positive. To test this we used threshold levels ranging from 0.1 to 0.5 and plotted the precision, recall and f1-score, as seen in Figure 5.20. As we can see the optimal f1-score is reached at a threshold of 0.3.

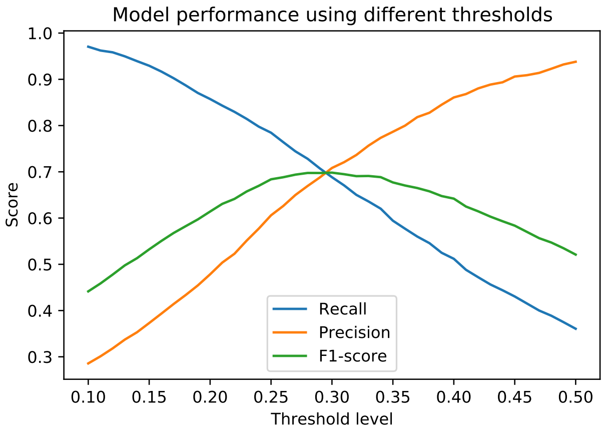

Figure 5.20: Model 1 performance using various threshold levels

The model performance using this threshold is shown in Table 5.26 and is compared to the model when using a threshold of 0.5. We can see that the model strikes a better balance between precision and recall. As with the two model previous models also this model's applicability depends on the needs in a specific real world setting.

	Precision	Recall	Precision diff	Recall diff
Class 0	0.93	0.94	+0.05	-0.06
Class 1	0.71	0.69	-0.23	+0.33

Table 5.26: Model performance and comparison when using a lower threshold

5.8 Overview

In Table 5.27 below we can see the performance scores of our models. The overall best performing model was using original data including age, gender and the heart related diagnosis flag. This model was also used when comparing threshold levels and showed the best balance in performance metrics as well as precision and recall balance when using a lower threshold.

			Performance scores			
		Data	AUROC	pre-rec (class 0)	pre-rec (class 1)	OOB
Base Model	Default	Original	0.77	0.93	0.49	0.84
	Optimized	Original	0.78	0.93	0.51	0.84
Balancing data	Undersample	RUS	0.76	0.97	0.69	0.68
		Near Miss	0.54	0.86	0.19	0.86
	Oversample	ROS	0.78	0.93	0.52	**0.98**
		SMOTE	0.76	0.93	0.47	0.88
	Class weights	Original	0.78	0.93	0.52	0.84
Patient-specific model	Gender	Original	0.79	0.94	0.54	0.85
		RUS	0.77	0.93	0.48	0.69
	Age	Original	0.88	0.97	0.70	0.87
		RUS	0.85	0.96	0.62	0.76
	Gender & age	Original	0.90	0.97	0.74	0.88
		RUS	0.86	0.96	0.65	0.77
Diagnoses-specific model	Heart related diagnosis subset	Original subset	0.91	0.97	**0.78**	0.89
		RUS subset	0.88	0.96	0.71	0.80
	Non heart related diagnosis subset	Original subset	0.88	**0.98**	**0.78**	0.89
		RUS subset	0.88	0.97	0.69	0.80
	(Non) heart related diagnosis variable	Original	**0.92**	**0.98**	**0.78**	0.88
		RUS	0.88	0.97	0.68	0.78

Table 5.27: Overview of model performances

Chapter 6

Future work

One of the key challenges in this thesis was to make a model that would predict deterioration using just the ECG signal and without knowing too much about the patient. Some basic patient information as gender, age and, if known, is the diagnosis was heart related were added to improve model performance. In future work this can be further extended to include more variables like for instance a patient's weight. Further optimizing the model's performance while keeping the requirements for use low by not asking hard to retrieve information as input will make the model more applicable for real world use.

For now the two models serve two different purposes; one with a high recall, aiming not to miss any deteriorating patients, and one with a high precision, aiming no to misclassify any recovering patients. Closer work with healthcare providers like the ICU in a hospital can provide more insights into what is most desired for each setting. This feedback can than be used to optimize a model more to their specific needs.

Another aspect that can be modified for real world use are the time frames used. All current tests are done using the last 24 hours of data, split in data per hour. This time window can be greater or smaller, as well as the time per instance, e.g. 48 hours of data, or 6 hours split per 30 minutes. Using a small time frame could for instance be used in cases where the model would be applied as more of a last minute warning system for cases that have been missed, where as a longer time frame could be used a more subtle supporting tool to confirm or contradict the healthcare worker's estimation of a patient.

Not only the time frames but also what features to extract from the ECG data can be altered. There was a clear difference in feature importance as shown in section 5.6. Excluding some features could decrease model complexity while not hurting performance much, which could be preferable in some applications.

Other data as from a difference dataset or collected personally could also allow for the extraction of frequency based ECG features such as peak heights, as this was not possible with the MIMIC dataset. These features could give way more insight and result in a higher performance.

As is the model only predicts if an instance of an hour belongs to someone who is deteriorating or recovering. As see in section 5.5 there is no clear trend in patient prediction closer to the end of stay. Such a trend would however be an even more helpful tool as it would take away much uncertainty that can arise from a score that changes each hour. If no clear trend is seen the model would only alert if it notices something, which could be temporary and only happen in a certain hour. With a trend over time however the change per hour could be taken into account and the direction of the trend could be used as an indicator for a patient's well being.

Aside from the differences desired model performance and real world applicability, the targeted population is also important. The original goal of this thesis was to predict deterioration in newborns in the NICU. If data on this population can be gathered, similar research can be done on this population.

Chapter 7

Conclusion

The aim of this thesis was to create a predictive model using a minimal amount of needed input, reaching an as high as possible performance. To tackle the first challenge we have used only data retrieved from one ECG lead in combination with the most basic patient related variables, namely age and gender. When the diagnosis is known to be heart related or not this can add to the performance. To reach the maximum result, separate models are needed for these two cases, as using the diagnosis as a separate input variable resulted in slightly worse performance. If this is favorable relies on the technical limitation of the application the models would be used in. The second part, the model performance, did not reach a height that would make it directly applicable in most settings. For now, either a lot of deteriorating patients are missed, making the model less useful when it comes to triggering early intervention, or the model classifies too many false positives, making the model less useful for resource management. The amount of false positives are however only high because of the imbalance in the data, which could be an accurate reflection of the real world. But this data is also collected from an American hospital, which could have a higher recovery rate than hospitals in low resource settings. Depending on the current situation and needs of a hospital in a low resource setting, the addition of either of these models could therefore still have a positive effect, although not optimal. Important to note is that in any scenario such a model would be used to assist healthcare workers, and not to take over their task of estimating a patient's well being or if he or she is showing signs of recovering or not. When adjusting the threshold we reach a model performance that could be considered high enough to be applied in a real world setting, especially taking into account the vast room for improvement in the hospitals in low resource settings.

The most challenging problem, besides only using ECG data, was the fact that the population was extremely diverse, considering a collection of patients in a wide age group, from all different backgrounds and different diagnoses. Some patients had multiple ICU stays, co-morbidities or other factors that could affect their state, while others might have only had a short stay and would be considered to fall in a completely different category. Fitting and evaluating a model on such a diverse group of people while barely taking into account patient specific variables deemed to be a big challenge and is most probably the reason for the less than optimal performance.

A lot of insights have however be gathered from these models such as what features are most important when making a prediction. A followup study with the same aim as ours can learn from the findings in this thesis and the models can be used side-by-side with newly trained ones for comparison in a real world environment.

Bibliography

[1] Python Lib tree interpreter. https://pypi.org/project/treeinterpreter/. Accessed: 2021-08-18. 40

[2] *Artificial Intelligence (AI) in Healthcare Market 2020-2026.* Orion Market Research Private Limited, 2021. 1

[3] U Rajendra Acharya, K Paul Joseph, Natarajan Kannathal, Choo Min Lim, and Jasjit S Suri. Heart rate variability: a review. *Medical and biological engineering and computing*, 44(12):1031–1051, 2006. 6

[4] Kevin Addison, MP Griffin, JR Moorman, DE Lake, and TM O'shea. Heart rate characteristics and neurodevelopmental outcome in very low birth weight infants. *Journal of Perinatology*, 29(11):750–756, 2009. 6

[5] S Arora, I Lang, V Nayyar, E Stachowski, and DL Ross. Atrial fibrillation in a tertiary care multidisciplinary intensive care unit—incidence and risk factors. *Anaesthesia and intensive care*, 35(5):707–713, 2007. 17

[6] Lisa M Askie, Brian A Darlow, Neil Finer, Barbara Schmidt, Ben Stenson, William Tarnow-Mordi, Peter G Davis, Waldemar A Carlo, Peter Brocklehurst, Lucy C Davies, et al. Association between oxygen saturation targeting and death or disability in extremely preterm infants in the neonatal oxygenation prospective meta-analysis collaboration. *Jama*, 319(21):2190–2201, 2018. 6

[7] Henrietta S Bada, Sheldon B Korones, Edward H Perry, Kristopher L Arheart, John D Ray, Massroor Pourcyrous, H Lynn Magill, William Runyan III, Grant W Somes, Frank C Clark, et al. Mean arterial blood pressure changes in premature infants and those at risk for intraventricular hemorrhage. *The Journal of pediatrics*, 117(4):607–614, 1990. 6

[8] Douglas P Barnaby, Shannon M Fernando, Christophe L Herry, Nathan B Scales, Edward John Gallagher, and Andrew JE Seely. Heart rate variability, clinical and laboratory measures to predict future deterioration in patients presenting with sepsis. *Shock*, 51(4):416–422, 2019. 8

[9] Ruchie Bhardwaj, Adhiraaj Sethi, and Raghunath Nambiar. Big data in genomics: An overview. In *2014 IEEE International Conference on Big Data (Big Data)*, pages 45–49. IEEE, 2014. 5

[10] Timothy W Bickmore, Laura M Pfeifer, and Brian W Jack. Taking the time to care: empowering low health literacy hospital patients with virtual nurse agents. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1265–1274, 2009. 2

[11] Guthrie S Birkhead, Michael Klompas, and Nirav R Shah. Uses of electronic health records for public health surveillance to advance public health. *Annual review of public health*, 36:345–359, 2015. 5

[12] MJ Bizzarro, FY Li, K Katz, V Shabanova, RA Ehrenkranz, and V Bhandari. Temporal quantification of oxygen saturation ranges: an effort to reduce hyperoxia in the neonatal intensive care unit. *Journal of Perinatology*, 34(1):33–38, 2014. 6

[13] Yu-Hsiang Chou, Wei-Lieh Huang, Chin-Hao Chang, Cheryl CH Yang, Terry BJ Kuo, Shuei-Liong Lin, Wen-Chih Chiang, and Tzong-Shinn Chu. Heart rate variability as a predictor of rapid renal function deterioration in chronic kidney disease patients. *Nephrology*, 24(8):806–813, 2019. 8

[14] F Cockburn, RWI Cooke, HR Gamsu, A Greenough, A Hopkins, N Mcintosh, SA Ogston, GJ Parry, M Silverman, JCL Shaw, et al. The crib (clinical risk index for babies) score: a tool for assessing initial neonatal risk and comparing performance of neonatal intensive care units. *The Lancet*, 342(8865):193–198, 1993. 6

[15] Steven Cunningham, Andrew G Symon, Robert A Elton, Changqing Zhu, and Neil McIntosh. Intra-arterial blood pressure reference ranges, death and morbidity in very low birthweight infants during the first seven days of life. *Early human development*, 56(2-3):151–165, 1999. 6

[16] A Das, M Mhanna, J Sears, JW Houdek, N Kumar, D Gunzler, D Einstadter, and M Collin. Effect of fluctuation of oxygenation and time spent in the target range on retinopathy of prematurity in extremely low birth weight infants. *Journal of neonatal-perinatal medicine*, 11(3):257–263, 2018. 6

[17] Thomas Davenport and Ravi Kalakota. The potential for artificial intelligence in healthcare. *Future healthcare journal*, 6(2):94, 2019. 1

[18] Juliann M Di Fiore, Farhad Kaffashi, Kenneth Loparo, Abdus Sattar, Mark Schluchter, Ryan Foglyano, Richard J Martin, and Christopher G Wilson. The relationship between patterns of intermittent hypoxia and retinopathy of prematurity in preterm infants. *Pediatric research*, 72(6):606–612, 2012. 6

[19] Juliann M Di Fiore, Richard J Martin, Hong Li, Nathan Morris, Waldemar A Carlo, Neil Finer, Michele Walsh, Alan H Jobe, Michael S Caplan, Richard A Polin, et al. Patterns of oxygenation, mortality, and growth status in the surfactant positive pressure and oxygen trial cohort. *The Journal of pediatrics*, 186:49–56, 2017. 6

[20] Kim Kopenhaver Doheny, C Palmer, Kirsteen N Browning, Puneet Jairath, Duanping Liao, Fan He, and RA Travagli. Diminished vagal tone is a predictive biomarker of necrotizing enterocolitis-risk in preterm infants. *Neurogastroenterology & Motility*, 26(6):832–840, 2014. 6

[21] Karen Fairchild, Mary Mohr, Alix Paget-Brown, Christa Tabacaru, Douglas Lake, John Delos, Joseph Randall Moorman, and John Kattwinkel. Clinical associations of immature breathing in preterm infants: part 1—central apnea. *Pediatric research*, 80(1):21–27, 2016. 6

[22] Karen D Fairchild. Predictive monitoring for early detection of sepsis in neonatal icu patients. *Current opinion in pediatrics*, 25(2):172–179, 2013. 7

[23] Karen D Fairchild and Judy L Aschner. Hero monitoring to reduce mortality in nicu patients. *Research and Reports in Neonatology*, 2:65–76, 2012. 7

[24] Karen D Fairchild and Douglas E Lake. Cross-correlation of heart rate and oxygen saturation in very low birthweight infants: association with apnea and adverse events. *American journal of perinatology*, 35(05):463–469, 2018. 6

[25] Karen D Fairchild, Douglas E Lake, John Kattwinkel, J Randall Moorman, David A Bateman, Philip G Grieve, Joseph R Isler, and Rakesh Sahni. Vital signs and their cross-correlation in sepsis and nec: a study of 1,065 very-low-birth-weight infants in two nicus. *Pediatric research*, 81(2):315–321, 2017. 6

[26] Karen D Fairchild, Robert L Schelonka, David A Kaufman, Waldemar A Carlo, John Kattwinkel, Peter J Porcelli, Cristina T Navarrete, Eduardo Bancalari, Judy L Aschner, M Whit Walker, et al. Septicemia mortality reduction in neonates in a heart rate characteristics monitoring trial. *Pediatric research*, 74(5):570–575, 2013. 6

[27] KD Fairchild, RA Sinkin, F Davalian, AE Blackman, JR Swanson, JA Matsumoto, DE Lake, JR Moorman, and JA Blackman. Abnormal heart rate characteristics are associated with abnormal neuroimaging and outcomes in extremely low birth weight infants. *Journal of Perinatology*, 34(5):375–379, 2014. 6

[28] Heladia García, Raúl Villegas-Silva, Dina Villanueva-García, Héctor González-Cabello, Marina López-Padilla, Arturo Fajardo-Gutiérrez, María del Carmen Martínez-García, and Juan Garduño-Espinosa. Validation of a prognostic index in the critically ill newborn. *PRISM (Paediatric Risk of Mortality)*, 14:16, 2000. 6

[29] Nitin Goel, Mallinath Chakraborty, William John Watkins, and Sujoy Banerjee. Predicting extubation outcomes—a model incorporating heart rate characteristics index. *The Journal of pediatrics*, 195:53–58, 2018. 6

[30] Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220, 2000. 3

[31] Ricki F Goldstein, Robert J Thompson, Jerri M Oehler, and Jane E Brazy. Influence of acidosis, hypoxemia, and hypotension on neurodevelopmental outcome in very low birth weight infants. *Pediatrics*, 95(2):238–243, 1995. 6

[32] James E Gray, Douglas K Richardson, Marie C McCormick, Kathryn Workman-Daniels, and Donald A Goldmann. Neonatal therapeutic intervention scoring system: a therapy-based severity-of-illness index. *Pediatrics*, 90(4):561–567, 1992. 6

[33] M Pamela Griffin, Douglas E Lake, T Michael O'Shea, and J Randall Moorman. Heart rate characteristics and clinical signs in neonatal sepsis. *Pediatric research*, 61(2):222–227, 2007. 6

[34] M Pamela Griffin, T Michael O'Shea, Eric A Bissonette, Frank E Harrell, Douglas E Lake, and J Randall Moorman. Abnormal heart rate characteristics preceding neonatal sepsis and sepsis-like illness. *Pediatric research*, 53(6):920–926, 2003. 6

[35] Przemyslaw Guzik, Jaroslaw Piskorski, Tomasz Krauze, Andrzej Wykretowicz, and Henryk Wysocki. Heart rate asymmetry by poincaré plots of rr intervals. 2006. 8

[36] Katharine E Henry, David N Hager, Peter J Pronovost, and Suchi Saria. A targeted real-time early warning score (trewscore) for septic shock. *Science translational medicine*, 7(299):299ra122–299ra122, 2015. 6

[37] Suma B Hoffman, Yun-Ju Cheng, Laurence S Magder, Narendra Shet, and Rose M Viscardi. Cerebral autoregulation in premature infants during the first 96 hours of life and relationship to adverse outcomes. *Archives of Disease in Childhood-Fetal and Neonatal Edition*, 104(5):F473–F479, 2019. 6

[38] AO Hofstetter, L Legnevall, E Herlenius, and M Katz-Salamon. Cardiorespiratory development in extremely preterm infants: vulnerability to infection and persistence of events beyond term-equivalent age. *Acta Paediatrica*, 97(3):285–292, 2008. 6

[39] JEFFREY D Horbar, LYNN Onstad, ELIZABETH Wright, Sumner J Yaffe, Charlotte Catz, LL Wright, MH Malloy, GG Rhoades, T Gordon, E Phillips, et al. Predicting mortality risk for infants weighing 501 to 1500 grams at birth: a national institutes of health neonatal research network report. *Critical care medicine*, 21(1):12–18, 1993. 6

[40] Nianzong Hou, Mingzhe Li, Lu He, Bing Xie, Lin Wang, Rumin Zhang, Yong Yu, Xiaodong Sun, Zhengsheng Pan, and Kai Wang. Predicting 30-days mortality for mimic-iii patients with sepsis-3: a machine learning approach using xgboost. *Journal of translational medicine*, 18(1):1–14, 2020. 7

[41] Peter B Jensen, Lars J Jensen, and Søren Brunak. Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics*, 13(6):395–405, 2012. 5

[42] Min Jiang, Yukun Chen, Mei Liu, S Trent Rosenbloom, Subramani Mani, Joshua C Denny, and Hua Xu. A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. *Journal of the American Medical Informatics Association*, 18(5):601–606, 2011. 5

[43] Alistair EW Johnson, Tom J Pollard, and Roger G Mark. Reproducibility in critical care: a mortality prediction case study. In *Machine Learning for Healthcare Conference*, pages 361–376. PMLR, 2017. 7

[44] Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, H. Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016. 3

[45] Pollard T. Johnson A. and Mark R. Mimic-iii clinical database (version 1.4). 2016. 3

[46] Rohan Joshi, Deedee Kommers, Laurien Oosterwijk, Loe Feijs, Carola Van Pul, and Peter Andriessen. Predicting neonatal sepsis using features of heart rate variability, respiratory characteristics, and ecg-derived estimates of infant motion. *IEEE journal of biomedical and health informatics*, 24(3):681–692, 2019. 6, 8

[47] Christoph U Lehmann, Karen G O'Connor, Vanessa A Shorte, and Timothy D Johnson. Use of electronic health record systems by office-based pediatricians. *Pediatrics*, 135(1):e7–e15, 2015. 5

[48] Guillaume Lemaître, Fernando Nogueira, and Christos K. Aridas. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18(17):1–5, 2017. 28

[49] JA Low, AB Froese, RS Galbraith, JT Smith, EE Sauerbrei, and EJ Derrick. The association between preterm newborn hypotension and hypoxemia and outcome during the first year. *Acta Paediatrica*, 82(5):433–437, 1993. 6

[50] Kemi K Mascoll-Robertson, Rose M Viscardi, and Hyung C Woo. The objective use of pulse oximetry to predict respiratory support transition in preterm infants: an observational pilot study. *Respiratory care*, 61(4):416–422, 2016. 6

[51] Scott Mayer McKinney, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, Trevor Back, Mary Chesus, Greg S Corrado, Ara Darzi, et al. International evaluation of an ai system for breast cancer screening. *Nature*, 577(7788):89–94, 2020. 1

[52] Vivienne M Miall-Allen, Linda S de Vries, Lilly MS Dubowitz, and Andrew GL Whitelaw. Blood pressure fluctuation and intraventricular hemorrhage in the preterm infant of less than 31 weeks' gestation. *Pediatrics*, 83(5):657–661, 1989. 6

[53] VM Miall-Allen, LS De Vries, and AG Whitelaw. Mean arterial blood pressure and neonatal cerebral lesions. *Archives of disease in childhood*, 62(10):1068–1069, 1987. 6

[54] MIT-LCP. Native python wfdb package. `https://github.com/MIT-LCP/wfdb-python`, 2021. 15

[55] Mary A Mohr, Karen D Fairchild, Manisha Patel, Robert A Sinkin, Matthew T Clark, J Randall Moorman, Douglas E Lake, John Kattwinkel, and John B Delos. Quantification of periodic breathing in premature infants. *Physiological measurement*, 36(7):1415, 2015. 6

[56] Moody G. Moody B., Clifford G. Villarroel M., and Silva I. Mimic-iii waveform database matched subset (version 1.0). 2020. 3

[57] Moody G. Moody B., Clifford G. Villarroel M., and Silva I. Mimic-iii waveform database (version 1.0). 2020. 3

[58] J Randall Moorman, Douglas E Lake, and M Pamela Griffin. Heart rate characteristics monitoring for neonatal sepsis. *IEEE Transactions on Biomedical Engineering*, 53(1):126–132, 2005. 7

[59] Joseph Randall Moorman, Waldemar A Carlo, John Kattwinkel, Robert L Schelonka, Peter J Porcelli, Christina T Navarrete, Eduardo Bancalari, Judy L Aschner, Marshall Whit Walker, Jose A Perez, et al. Mortality reduction by heart rate characteristic monitoring in very low birth weight neonates: a randomized trial. *The Journal of pediatrics*, 159(6):900–906, 2011. 6

[60] M. T. Moss T. J., Clark, Enfield Calland J. F., Voss J. D. Lake K. B., D. E., and Moorman J. R. Cardiorespiratory dynamics measured from continuous ecg monitoring improves detection of deterioration in acute care patients: A retrospective cohort study. *PloS one*, 12(8):e0181448, 2017. 8

[61] Travis B Murdoch and Allan S Detsky. The inevitable application of big data to health care. *Jama*, 309(13):1351–1352, 2013. 5

[62] Saraladevi Naicker, Jacob Plange-Rhule, Roger C Tutt, and John B Eastwood. Shortage of healthcare workers in developing countries–africa. *Ethnicity & disease*, 19(1):60, 2009. 5

[63] Shane O'Sullivan, Nathalie Nevejans, Colin Allen, Andrew Blyth, Simon Leonard, Ugo Pagallo, Katharina Holzinger, Andreas Holzinger, Mohammed Imran Sajid, and Hutan Ashrafian. Legal, regulatory, and ethical frameworks for development of standards in artificial intelligence (ai) and autonomous robotic surgery. *The International Journal of Medical Robotics and Computer Assisted Surgery*, 15(1):e1968, 2019. 2

[64] Gareth Parry, Janet Tucker, William Tarnow-Mordi, UK Neonatal Staffing Study Collaborative Group, et al. Crib ii: an update of the clinical risk index for babies score. *The Lancet*, 361(9371):1789–1791, 2003. 6

[65] Manisha Patel, Mary Mohr, Douglas Lake, John Delos, J Randall Moorman, Robert A Sinkin, John Kattwinkel, and Karen Fairchild. Clinical associations with immature breathing in preterm infants: part 2—periodic breathing. *Pediatric research*, 80(1):28–34, 2016. 6

[66] Vimla L Patel, Edward H Shortliffe, Mario Stefanelli, Peter Szolovits, Michael R Berthold, Riccardo Bellazzi, and Ameen Abu-Hanna. The coming of age of artificial intelligence in medicine. *Artificial intelligence in medicine*, 46(1):5–17, 2009. 5

[67] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. 24

[68] Jeffrey M Perlman, Joseph B McMenamin, and Joseph J Volpe. Fluctuating cerebral blood-flow velocity in respiratory-distress syndrome: relation to the development of intraventricular hemorrhage. *New England Journal of Medicine*, 309(4):204–209, 1983. 6

[69] Christian F Poets, Robin S Roberts, Barbara Schmidt, Robin K Whyte, Elizabeth V Asztalos, David Bader, Aida Bairam, Diane Moddemann, Abraham Peliowski, Yacov Rabi, et al. Association between intermittent hypoxemia or bradycardia and late death or disability in extremely preterm infants. *Jama*, 314(6):595–603, 2015. 6

[70] Alberto Porta, S Guzzetti, N Montano, T Gnecchi-Ruscone, R Furlan, and A Malliani. Time reversibility in short-term heart period variability. In *2006 Computers in Cardiology*, pages 77–80. IEEE, 2006. 9

[71] Amro Qaddoura, Geneviève C Digby, Conrad Kabali, Piotr Kukla, Zhong-Qun Zhan, and Adrian M Baranchuk. The value of electrocardiography in prognosticating clinical deterioration and mortality in acute pulmonary embolism: A systematic review and meta-analysis. *Clinical cardiology*, 40(10):814–824, 2017. 8

[72] Thomas M Raffay, Andrew M Dylag, Abdus Sattar, Elie G Abu Jawdeh, Shufen Cao, Benjamin M Pax, Kenneth A Loparo, Richard J Martin, and Juliann M Di Fiore. Neonatal intermittent hypoxemia events are associated with diagnosis of bronchopulmonary dysplasia at 36 weeks postmenstrual age. *Pediatric research*, 85(3):318–323, 2019. 6

[73] G Rauscher, AC DeGiorgio, PR Miller, and CM DeGiorgio. Sudden unexpected death in epilepsy associated with progressive deterioration in heart rate variability. *Epilepsy & Behavior*, 21(1):103–105, 2011. 8

[74] Dougbas K Richardson, Ciaran S Phibbs, James E Gray, Marie C McCormick, Kathryn Workman-Daniels, and Donald A Goldmann. Birth weight and illness severity: independent predictors of neonatal mortality. *Pediatrics*, 91(5):969–975, 1993. 6

[75] Douglas K Richardson, John D Corcoran, Gabriel J Escobar, Shoo K Lee, et al. Snap-ii and snappe-ii: simplified newborn illness severity and mortality risk scores. *The Journal of pediatrics*, 138(1):92–100, 2001. 6

[76] Douglas K Richardson, James E Gray, Marie C McCormick, Kathryn Workman, and Donald A Goldmann. Score for neonatal acute physiology: a physiologic severity index for neonatal intensive care. *Pediatrics*, 91(3):617–623, 1993. 6

[77] James Rickert. Patient-centered care: what it means and how to get there. *Health Affairs Blog*, 24:1–4, 2012. 5

[78] Oksana Semenova, Gordon Lightbody, John M O'Toole, Geraldine Boylan, Eugene Dempsey, and Andriy Temko. Coupling between mean blood pressure and eeg in preterm neonates is associated with reduced illness severity scores. *PloS one*, 13(6):e0199587, 2018. 6

[79] Janet S Soul, Peter E Hammer, Miles Tsuji, J Philip Saul, Haim Bassan, Catherine Limperopoulos, Donald N Disalvo, Marianne Moore, Patricia Akins, Steven Ringer, et al. Fluctuating pressure-passivity is common in the cerebral circulation of sick premature infants. *Pediatric research*, 61(4):467–473, 2007. 6

[80] Matthew L Stone, Philip M Tatum, Jörn-Hendrik Weitkamp, Anamika B Mukherjee, Joshua Attridge, Eugene D McGahren, Bradley M Rodgers, Douglas E Lake, J Randall Moorman, and Karen D Fairchild. Abnormal heart rate characteristics before clinical diagnosis of necrotizing enterocolitis. *Journal of Perinatology*, 33(11):847–850, 2013. 6

[81] BA Sullivan, A Wallman-Stokes, J Isler, R Sahni, JR Moorman, KD Fairchild, and DE Lake. Early pulse oximetry data improves prediction of death and adverse outcomes in a two-center cohort of very low birth weight infants. *American journal of perinatology*, 35(13):1331–1338, 2018. 6

[82] Brynne A Sullivan, Stephanie M Grice, Douglas E Lake, J Randall Moorman, and Karen D Fairchild. Infection and other clinical correlates of abnormal heart rate characteristics in preterm infants. *The Journal of pediatrics*, 164(4):775–780, 2014. 6

[83] Brynne A Sullivan, Christina McClure, Jamie Hicks, Douglas E Lake, J Randall Moorman, and Karen D Fairchild. Early heart rate characteristics predict death and morbidities in preterm infants. *The Journal of pediatrics*, 174:57–62, 2016. 6

[84] Christa R Tabacaru, Suk Young Jang, Manisha Patel, Faranek Davalian, Santina Zanelli, and Karen D Fairchild. Impact of caffeine boluses and caffeine discontinuation on apnea and hypoxemia in preterm infants. *Journal of caffeine research*, 7(3):103–110, 2017. 6

[85] Aline Taoum, Farah Mourad-Chehade, Hassan Amoud, and Ziad Fawal. Predicting ards using the mimic ii physiological database. In *2016 IEEE International Multidisciplinary Conference on Engineering Technology (IMCET)*, pages 47–51. IEEE, 2016. 7

[86] Brooke D Vergales, Santina A Zanelli, Julie A Matsumoto, Howard P Goodkin, Douglas E Lake, J Randall Moorman, and Karen D Fairchild. Depressed heart rate variability is associated with abnormal eeg, mri, and death in neonates with hypoxic ischemic encephalopathy. *American journal of perinatology*, 31(10):855–862, 2014. 6

[87] Zachary A Vesoulis, Rachel L Bank, Doug Lake, Aaron Wallman-Stokes, Rakesh Sahni, J Randall Moorman, Joseph R Isler, Karen D Fairchild, and Amit M Mathur. Early hypoxemia burden is strongly associated with severe intracranial hemorrhage in preterm infants. *Journal of Perinatology*, 39(1):48–53, 2019. 6

[88] Alyssa Warburton, Ranjan Monga, Venkatesh Sampath, and Navin Kumar. Continuous pulse oximetry and respiratory rate trends predict short-term respiratory and growth outcomes in premature infants. *Pediatric research*, 85(4):494–501, 2019. 6

[89] Chris Watts and Ijeoma Ibegbulam. Access to electronic healthcare information resources in developing countries: Experiences from the medical library, college of medicine, university of nigeria. *IFLA journal*, 32(1):54–61, 2006. 5

[90] Peter H Westfall. Kurtosis as peakedness, 1905–2014. rip. *The American Statistician*, 68(3):191–195, 2014. 19

[91] Xian Zeng, Gang Yu, Yang Lu, Linhua Tan, Xiujing Wu, Shanshan Shi, Huilong Duan, Qiang Shu, and Haomin Li. Pic, a paediatric-specific intensive care database. *Scientific data*, 7(1):1–8, 2020. 3

Appendix A

Appendix

A.1 Example header file

```
3544749_0001 4 125 3811 17:48:00.230
3544749_0001.dat 80 1/mV 8 0 -72 4808 0 II
3544749_0001.dat 80 1/mV 8 0 -72 -14144 0 AVF
3544749_0001.dat 80 1/mmHg 8 0 -72 8708 0 ABP
3544749_0001.dat 80 1/mmHg 8 0 -72 4808 0 PAP
```

A.2 Filtered IDC9 Codes

ROW_ID	ICD9_CODE	SHORT_TITLE	LONG_TITLE
369	0860	Chagas disease of heart	Chagas' disease with heart involvement
423	09389	Cardiovascular syph NEC	Other specified cardiovascular syphilis
488	09885	Gonococcal heart dis NEC	Other gonococcal heart disease
1332	1641	Malignant neopl heart	Malignant neoplasm of heart
2423	2127	Benign neoplasm heart	Benign neoplasm of heart
4299	39890	Rheumatic heart dis NOS	Rheumatic heart disease, unspecified
4300	39891	Rheumatic heart failure	Rheumatic heart failure (congestive)
4301	39899	Rheumatic heart dis NEC	Other rheumatic heart diseases
4306	40201	Mal hypert hrt dis w hf	Malignant hypertensive heart disease with heart failure
4308	40211	Benign hyp ht dis w hf	Benign hypertensive heart disease with heart failure
4310	40291	Hyp ht dis NOS w ht fail	Unspecified hypertensive heart disease with heart failure
4318	40401	Mal hyp ht/kd I-IV w hf	Hypertensive heart and chronic kidney disease, malignant, with heart failure and with chronic kidney disease stage I through stage IV, or unspecified
4320	40403	Mal hyp ht/kd stg V w hf	Hypertensive heart and chronic kidney disease, malignant, with heart failure and with chronic kidney disease stage V or end stage renal disease
4322	40411	Ben hyp ht/kd I-IV w hf	Hypertensive heart and chronic kidney disease, benign, with heart failure and with chronic kidney disease stage I through stage IV, or unspecified
4324	40413	Ben hyp ht/kd stg V w hf	Hypertensive heart and chronic kidney disease, benign, with heart failure and chronic kidney disease stage V or end stage renal disease
4326	40491	Hyp ht/kd NOS I-IV w hf	Hypertensive heart and chronic kidney disease, unspecified, with heart failure and with chronic kidney disease stage I through stage IV, or unspecified
4328	40493	Hyp ht/kd NOS st V w hf	Hypertensive heart and chronic kidney disease, unspecified, with heart failure and chronic kidney disease stage V or end stage renal disease
4335	41000	AMI anterolateral,unspec	Acute myocardial infarction of anterolateral wall, episode of care unspecified
4336	41001	AMI anterolateral, init	Acute myocardial infarction of anterolateral wall, initial episode of care
4337	41002	AMI anterolateral,subseq	Acute myocardial infarction of anterolateral wall, subsequent episode of care
4338	41010	AMI anterior wall,unspec	Acute myocardial infarction of other anterior wall, episode of care unspecified
4339	41011	AMI anterior wall, init	Acute myocardial infarction of other anterior wall, initial episode of care
4340	41012	AMI anterior wall,subseq	Acute myocardial infarction of other anterior wall, subsequent episode of care
4341	41020	AMI inferolateral,unspec	Acute myocardial infarction of inferolateral wall, episode of care unspecified
4342	41021	AMI inferolateral, init	Acute myocardial infarction of inferolateral wall, initial episode of care
4343	41022	AMI inferolateral,subseq	Acute myocardial infarction of inferolateral wall, subsequent episode of care

4344	41030	AMI inferopost, unspec	Acute myocardial infarction of inferoposterior wall, episode of care unspecified
4345	41031	AMI inferopost, initial	Acute myocardial infarction of inferoposterior wall, initial episode of care
4346	41032	AMI inferopost, subseq	Acute myocardial infarction of inferoposterior wall, subsequent episode of care
4347	41040	AMI inferior wall,unspec	Acute myocardial infarction of other inferior wall, episode of care unspecified
4348	41041	AMI inferior wall, init	Acute myocardial infarction of other inferior wall, initial episode of care
4349	41042	AMI inferior wall,subseq	Acute myocardial infarction of other inferior wall, subsequent episode of care
4350	41050	AMI lateral NEC, unspec	Acute myocardial infarction of other lateral wall, episode of care unspecified
4351	41051	AMI lateral NEC, initial	Acute myocardial infarction of other lateral wall, initial episode of care
4352	41052	AMI lateral NEC, subseq	Acute myocardial infarction of other lateral wall, subsequent episode of care
4359	41080	AMI NEC, unspecified	Acute myocardial infarction of other specified sites, episode of care unspecified
4360	41081	AMI NEC, initial	Acute myocardial infarction of other specified sites, initial episode of care
4361	41082	AMI NEC, subsequent	Acute myocardial infarction of other specified sites, subsequent episode of care
4362	41090	AMI NOS, unspecified	Acute myocardial infarction of unspecified site, episode of care unspecified
4363	41091	AMI NOS, initial	Acute myocardial infarction of unspecified site, initial episode of care
4364	41092	AMI NOS, subsequent	Acute myocardial infarction of unspecified site, subsequent episode of care
4365	4110	Post MI syndrome	Postmyocardial infarction syndrome
4366	4111	Intermed coronary synd	Intermediate coronary syndrome
4367	41181	Acute cor occlsn w/o MI	Acute coronary occlusion without myocardial infarction
4368	41189	Ac ischemic hrt dis NEC	Other acute and subacute forms of ischemic heart disease, other
4369	412	Old myocardial infarct	Old myocardial infarction
4374	41401	Crnry athrscl natve vssl	Coronary atherosclerosis of native coronary artery
4379	41406	Cor ath natv art tp hrt	Coronary atherosclerosis of native coronary artery of transplanted heart
4380	41407	Cor ath bps graft tp hrt	Coronary atherosclerosis of bypass graft (artery) (vein) of transplanted heart
4381	41410	Aneurysm of heart	Aneurysm of heart (wall)
4382	41411	Aneurysm coronary vessel	Aneurysm of coronary vessels
4383	41412	Dissection cor artery	Dissection of coronary artery
4384	41419	Aneurysm of heart NEC	Other aneurysm of heart
4385	4142	Chr tot occlus cor artry	Chronic total occlusion of coronary artery
4387	4144	Cor ath d/t calc cor lsn	Coronary atherosclerosis due to calcified coronary lesion
4388	4148	Chr ischemic hrt dis NEC	Other specified forms of chronic ischemic heart disease
4389	4149	Chr ischemic hrt dis NOS	Chronic ischemic heart disease, unspecified
4396	4161	Kyphoscoliotic heart dis	Kyphoscoliotic heart disease
4398	4168	Chr pulmon heart dis NEC	Other chronic pulmonary heart diseases
4399	4169	Chr pulmon heart dis NOS	Chronic pulmonary heart disease, unspecified
4430	4250	Endomyocardial fibrosis	Endomyocardial fibrosis

4440	4260	Atriovent block complete	Atrioventricular block, complete
4441	42610	Atriovent block NOS	Atrioventricular block, unspecified
4442	42611	Atriovent block-1st degr	First degree atrioventricular block
4443	42612	Atrioven block-mobitz ii	Mobitz (type) II atrioventricular block
4444	42613	Av block-2nd degree NEC	Other second degree atrioventricular block
4453	4266	Other heart block	Other heart block
4454	4267	Anomalous av excitation	Anomalous atrioventricular excitation
4459	4270	Parox atrial tachycardia	Paroxysmal supraventricular tachycardia
4460	4271	Parox ventric tachycard	Paroxysmal ventricular tachycardia
4468	42761	Atrial premature beats	Supraventricular premature beats
4470	42781	Sinoatrial node dysfunct	Sinoatrial node dysfunction
4471	42789	Cardiac dysrhythmias NEC	Other specified cardiac dysrhythmias
4473	4280	CHF NOS	Congestive heart failure, unspecified
4474	4281	Left heart failure	Left heart failure
4475	42820	Systolic hrt failure NOS	Systolic heart failure, unspecified
4476	42821	Ac systolic hrt failure	Acute systolic heart failure
4477	42822	Chr systolic hrt failure	Chronic systolic heart failure
4478	42823	Ac on chr syst hrt fail	Acute on chronic systolic heart failure
4479	42830	Diastolc hrt failure NOS	Diastolic heart failure, unspecified
4480	42831	Ac diastolic hrt failure	Acute diastolic heart failure
4481	42832	Chr diastolic hrt fail	Chronic diastolic heart failure
4482	42833	Ac on chr diast hrt fail	Acute on chronic diastolic heart failure
4483	42840	Syst/diast hrt fail NOS	Combined systolic and diastolic heart failure, unspecified
4484	42841	Ac syst/diastol hrt fail	Acute combined systolic and diastolic heart failure
4485	42842	Chr syst/diastl hrt fail	Chronic combined systolic and diastolic heart failure
4486	42843	Ac/chr syst/dia hrt fail	Acute on chronic combined systolic and diastolic heart failure
4492	4294	Hrt dis postcardiac surg	Functional disturbances following cardiac surgery
4495	42971	Acq cardiac septl defect	Acquired cardiac septal defect
4496	42979	Other sequelae of MI NEC	Certain sequelae of myocardial infarction, not elsewhere classified, other
4498	42982	Hyperkinetic heart dis	Hyperkinetic heart disease
4500	42989	Ill-defined hrt dis NEC	Other ill-defined heart diseases
4544	390	Rheum fev w/o hrt involv	Rheumatic fever without mention of heart involvement
4548	3918	Ac rheumat hrt dis NEC	Other acute rheumatic heart disease
4549	3919	Ac rheumat hrt dis NOS	Acute rheumatic heart disease, unspecified
4550	3920	Rheum chorea w hrt invol	Rheumatic chorea with heart involvement

4551	3929	Rheumatic chorea NOS	Rheumatic chorea without mention of heart involvement
6470	64850	Congen CV dis preg-unsp	Congenital cardiovascular disorders of mother, unspecified as to episode of care or not applicable
6471	64851	Congen CV dis-delivered	Congenital cardiovascular disorders of mother, delivered, with or without mention of antepartum condition
6472	64852	Congen CV dis-del w p/p	Congenital cardiovascular disorders of mother, delivered, with mention of postpartum complication
6473	64853	Congen CV dis-antepartum	Congenital cardiovascular disorders of mother, antepartum condition or complication
6474	64854	Congen CV dis-postpartum	Congenital cardiovascular disorders of mother, postpartum condition or complication
6475	64860	CV dis NEC preg-unspec	Other cardiovascular diseases of mother, unspecified as to episode of care or not applicable
6476	64861	CV dis NEC preg-deliver	Other cardiovascular diseases of mother, delivered, with or without mention of antepartum condition
6477	64862	CV dis NEC-deliver w p/p	Other cardiovascular diseases of mother, delivered, with mention of postpartum complication
6478	64863	CV dis NEC-antepartum	Other cardiovascular diseases of mother, antepartum condition or complication
6479	64864	CV dis NEC-postpartum	Other cardiovascular diseases of mother, postpartum condition or complication
7112	65970	Abn ftl hrt rate/rhy-uns	Abnormality in fetal heart rate or rhythm, unspecified as to episode of care or not applicable
7113	65971	Abn ftl hrt rate/rhy-del	Abnormality in fetal heart rate or rhythm, delivered, with or without mention of antepartum condition
7114	65973	Abn ftl hrt rate/rhy-ant	Abnormality in fetal heart rate or rhythm, antepartum condition or complication
7314	7455	Secundum atrial sept def	Ostium secundum type atrial septal defect
7319	7458	Septal closure anom NEC	Other bulbus cordis anomalies and anomalies of cardiac septal closure
7331	7467	Hypoplas left heart synd	Hypoplastic left heart syndrome
7335	74684	Obstruct heart anom NEC	Obstructive anomalies of heart, not elsewhere classified
7337	74686	Congenital heart block	Congenital heart block
7338	74687	Malposition of heart	Malposition of heart and cardiac apex
7339	74689	Cong heart anomaly NEC	Other specified congenital anomalies of heart
7340	7469	Cong heart anomaly NOS	Unspecified congenital anomaly of heart
7936	76381	Ab ftl hrt rt/rh b/f lab	Abnormality in fetal heart rate or rhythm before the onset of labor
7937	76382	Ab ftl hrt rt/rh dur lab	Abnormality in fetal heart rate or rhythm during labor
7938	76383	Ab ftl hrt rt/rhy NOS	Abnormality in fetal heart rate or rhythm, unspecified as to time of onset
8588	86102	Heart laceration-closed	Laceration of heart without penetration of heart chambers or without mention of open wound into thorax
8589	86103	Heart chamber lacerat-cl	Laceration of heart with penetration of heart chambers without mention of open wound into thorax
8590	86110	Heart injury NOS-open	Unspecified injury of heart with open wound into thorax

8591	86111	Heart contusion-open	Contusion of heart with open wound into thorax
8592	86112	Heart laceration-open	Laceration of heart without penetration of heart chambers, with open wound into thorax
8593	86113	Heart chamber lacer-opn	Laceration of heart with penetration of heart chambers with open wound into thorax
8586	86100	Heart injury NOS-closed	Unspecified injury of heart without mention of open wound into thorax
8587	86101	Heart contusion-closed	Contusion of heart without mention of open wound into thorax
8951	7797	Perivent leukomalacia	Periventricular leukomalacia
9055	77210	NB intraven hem NOS	Intraventricular hemorrhage unspecified grade
9056	77211	NB intraven hem,grade i	Intraventricular hemorrhage, grade I
9057	77212	NB intraven hem,grade ii	Intraventricular hemorrhage, grade II
9058	77213	NB intravn hem,grade iii	Intraventricular hemorrhage, grade III
9059	77214	NB intraven hem,grade iv	Intraventricular hemorrhage, grade IV
11948	V4321	Heart assist dev replace	Organ or tissue replaced by other means, heart assist device
11949	V4322	Artficial heart replace	Organ or tissue replaced by other means, fully implantable artificial heart
12276	9720	Pois-card rhythm regulat	Poisoning by cardiac rhythm regulators
12280	9724	Pois-coronary vasodilat	Poisoning by coronary vasodilators
12285	9729	Pois-cardiovasc agt NEC	Poisoning by other and unspecified agents primarily affecting the cardiovascular system
10140	V4581	Aortocoronary bypass	Aortocoronary bypass status
13385	E8726	Fail sterile heart cath	Failure of sterile precautions during heart catheterization
10141	V4582	Status-post ptca	Percutaneous transluminal coronary angioplasty status
10458	9920	Heat stroke & sunstroke	Heat stroke and sunstroke
9881	V4500	Status cardc dvce unspcf	Unspecified cardiac device in situ
9883	V4502	Status autm crd dfbrltr	Automatic implantable cardiac defibrillator in situ
9884	V4509	Status oth spcf crdc dvc	Other specified cardiac device in situ
11407	99600	Malfunc card dev/grf NOS	Mechanical complication of unspecified cardiac device, implant, and graft
11408	99601	Malfunc cardiac pacemake	Mechanical complication due to cardiac pacemaker (electrode)
11409	99602	Malfunc prosth hrt valve	Mechanical complication due to heart valve prosthesis
11410	99603	Malfunc coron bypass grf	Mechanical complication due to coronary bypass graft
11411	99604	Mch cmp autm mplnt dfbrl	Mechanical complication of automatic implantable cardiac defibrillator
11412	99609	Malfunc card dev/grf NEC	Other mechanical complication of cardiac device, implant, and graft
11437	99661	React-cardiac dev/graft	Infection and inflammatory reaction due to cardiac device, implant, and graft
12586	E8706	Acc cut/hem w heart cath	Accidental cut, puncture, perforation or hemorrhage during heart catheterization
12596	E8716	FB post heart catheter	Foreign object left in body during heart catheterization
11993	V5301	Adj cerebral vent shunt	Fitting and adjustment of cerebral ventricular (communicating) shunt
11998	V5331	Ftng cardiac pacemaker	Fitting and adjustment of cardiac pacemaker
11999	V5332	Ftng autmtc dfibrillator	Fitting and adjustment of automatic implantable cardiac defibrillator

12000	V5339	Ftng oth cardiac device	Fitting and adjustment of other cardiac device
14501	V810	Scrn-ischemic heart dis	Screening for ischemic heart disease
14503	V812	Screen-cardiovasc NEC	Screening for other and unspecified cardiovascular conditions
12346	99683	Compl heart transplant	Complications of transplanted heart
13740	E9429	Adv eff cardiovasc NEC	Other and unspecified agents primarily affecting the cardiovascular system causing adverse effects in therapeutic use
10949	79430	Abn cardiovasc study NOS	Abnormal cardiovascular function study, unspecified
10951	79439	Abn cardiovasc study NEC	Other nonspecific abnormal results of function study of cardiovascular system
12986	7852	Cardiac murmurs NEC	Undiagnosed cardiac murmurs
12987	7853	Abnorm heart sounds NEC	Other abnormal heart sounds
12994	7859	Cardiovas sys symp NEC	Other symptoms involving cardiovascular system
13947	V1253	Hx sudden cardiac arrest	Personal history of sudden cardiac arrest
9454	V1365	Hx-cong malform-heart	Personal history of (corrected) congenital malformations of heart and circulatory system
10078	V151	Hx-major cardiovasc surg	Personal history of surgery to heart and great vessels, presenting hazards to health
10120	V171	Family hx-stroke	Family history of stroke (cerebrovascular)
10122	V173	Fam hx-ischem heart dis	Family history of ischemic heart disease
10123	V1741	Fam hx sudden card death	Family history of sudden cardiac death (SCD)
10124	V1749	Fam hx-cardiovas dis NEC	Family history of other cardiovascular diseases
11613	E8583	Acc poisn-cardiovasc agt	Accidental poisoning by agents primarily affecting cardiovascular system
11630	99671	Comp-heart valve prosth	Other complications due to heart valve prosthesis
11631	99672	Comp-oth cardiac device	Other complications due to other cardiac device, implant, and graft
13402	E8745	Instrmnt fail-heart cath	Mechanical failure of instrument or apparatus during heart catheterization
14061	V717	Obs-susp cardiovasc dis	Observation for suspected cardiovascular disease
14084	V7281	Preop cardiovsclr exam	Pre-operative cardiovascular examination