Predictive Alarm Prevention by Forecasting Threshold Alarms at the Intensive Care Unit*

Jonas Chromik[1 [0000−0002−5709−4381]], Bjarne Pfitzner[1 [0000−0001−7824−8872]], Nina Ihde[1 [0000−0001−5776−3322]], Marius Michaelis[1 [0000−0002−6437−7152]], Denise Schmidt[1 [0000−0002−6299−0738]], Sophie Anne Ines Klopfenstein[2 [0000−0002−8470−2258]], Akira-Sebastian Poncette[2 [0000−0003−4627−7016]], Felix Balzer[2 [0000−0003−1575−2056]], and Bert Arnrich[1 [0000−0001−8380−7667]]

[1] Hasso Plattner Institute, University of Potsdam, Germany
{jonas.chromik, bjarne.pfitzner, bert.arnrich}@hpi.de
{nina.ihde, marius.michealis, denise.schmidt}@student.hpi.de
https://hpi.de/arnrich
[2] Charité – Universitätsmedizin Berlin, Germany
{sophie.klopfenstein, akira-sebastian.poncette, felix.balzer}@charite.de
https://medinfo.charite.de

Abstract. Patient monitors at intensive care units produce too many alarms – most of them being unnecessary. Medical staff becomes desensitised and ignores alarms. This phenomenon is called alarm fatigue and it negatively influences for both patients and staff. Some alarms are due to an acute and unforeseeable events but others are the result of a continued trend and hence foreseeable. We present a system that forecasts alarms – at least the foreseeable share – and transforms them into scheduled tasks. To achieve this, we use time-series models to forecast the patient's vital parameters and check whether the forecast violates the corresponding alarm threshold. The vital parameter measurements and alarm data stem from MIMIC-III but go through extensive preprocessing before the actual forecasting can take place. The result is a proof of concept but unfit for productive use. Lack of alarm data and low sampling frequencies for vital parameters impair alarm forecasting. Our work shows that gated recurrent unit models generally perform best for this task. A next step towards productive use is evaluating the approach on vital parameter data with higher time-resolution.

Keywords: Patient monitoring · Medical alarms · Alarm fatigue · Time-series forecasting

* This work was partially carried out within the INALO project. INALO is a cooperation project between AICURA medical GmbH, Charité – Universitätsmedizin Berlin, idalab GmbH, and Hasso Plattner Institute. INALO is funded by the German Federal Ministry of Education and Research under grant 16SV8559.

1 Introduction

Too many alarms from various devices make the intensive care unit a noisy and stressful place for both patients and medical staff. Different studies and reviews report numbers as high as 700 alarms per day [6] or 187 alarms per bed per day – only counting audible alarms [8]. Most of the alarms are either technically false, clinically irrelevant, or otherwise unnecessary [24,27]. This causes *alarm fatigue*: a desensitisation of clinicians by numerous alarms, many of which are either false or otherwise irrelevant [24,8]. Alarm fatigue negatively influences both for patients and medical staff [6].

The effects on medical staff are not extensively studied yet [23]. But studies indicate impaired mental efficiency and short-term memory [17] as well as stress [16] and stress-induced ailments such as burn-out [26]. A very recent study showed that false alarms are the medical staff's main concern and almost all members of the intensive care unit's staff requested a reduction in false alarms [22].

In patients, too many alarms and too much noise cause sleep deprivation [20], cardiovascular abnormalities [1,11], longer hospital stays [9], increased re-hospitalisation rates [11], increased need for analgesic medication [15], delayed wound healing [28], intensive care unit syndrome (a cluster of psychological and cognitive impairments) [2], and feelings of vulnerability and fear [10].

We still do not know how to effectively reduce the alarm burden on intensive care units and no single existing solution seems to be sufficient. In this work, we address threshold alarms – the most frequent type of alarms [8]. Threshold alarms on patient monitors inform the medical staff that a vital parameter – such as heart rate or blood pressure – is either too low or too high. If we could forecast these alarms, we could spot critical conditions early – even before they pose a danger to the patient. And we could convert acute, urgent, and disruptive alarms into scheduled tasks that the medical staff can handle at their own discretion – flexibly and over an extended period of time.

To forecast threshold alarms, we try to forecast the associated vital parameter. This is a regression task that we can solve with statistical or machine learning models for time-series. We try to find in the vital parameter measurements that will result in the vital parameter crossing one of the respective alarm thresholds. We are certain that we can not forecast all alarms since some alarms are actually the result of an acute event rather than a continued trend. But some alarms are foreseeable by inspecting the vital parameter measurements – these are the alarms we try to forecast and convert into scheduled tasks. We aim for few false positives to avoid increasing the staff's workload any further. But we are willing to accept many false negatives since not every alarm is foreseeable from the vital parameter trend.

This is an extended version of a conference paper [4]. We include content that we could not report in the conference paper due to page limitations, especially regarding data set preparation. The rest of this work is structured as follows: In section 2 we describe the data we use, why we chose this data set, and how we have to prepare it before usage. In section 3 we describe the methods we use for

alarm forecasting. In section 4 we show results on how well alarm forecasting works and which methods appear to be the most promising. Finally, in section 5 we discuss our results including limitations and directions for future work.

2 Data Preparation

We surveyed the variety of medical data sets and found several large intensive care unit data sets. We assessed the 3rd version of the Medical Information Mart for Intensive Care (MIMIC-III) [14], the eICU Collaborative Research Database (eICU CRD) [21], the High Time Resolution ICU Data Set (HiRID) [13], and the Amsterdam University Medical Centers Database (AmsterdamUMCdb) [25] and found that none of these data sets records patient monitor alarms. In fact, eICU CRD, HiRID, and AmsterdamUMCdb contain no alarm system information at all. MIMIC-III does not record alarm *events* but at least alarm *thresholds* and changes to these thresholds. Using thresholds and vital parameter measurements, we can reconstruct when alarms went off although the data set does not explicitly record these alarm events.

Other data sets record vital parameters with a higher temporal resolution (for example up to $f_s = 5\text{min}^{-1}$ for eICU CRD) and might allow for better results in the regression task described in section 3. But we definitively need the alarm threshold information – which are unique to MIMIC-III – to determine when an alarm goes off.

In the remainder of this section we describe how we reconstructed alarm events from the data provided in MIMIC-III. We used the methods described in [3] and we reproduce parts of these methods here.

2.1 Data Slicing

Data set preparation starts with reducing the data set to a subset of relevant information. MIMIC-III contains many table with various information such as diagnoses, lab values, and medications. But we are only interested in a single table: The CHARTEVENTS table records so called "charted events" – among them vital parameter measurements and alarm threshold updates. In CHARTEVENTS, all charted events are coded as data items with a unique ITEMID that identified the type of event, measurement, or value. The D_ITEMS table resolves the different ITEMIDs. We are only interested in the data items representing measurements for heart rate (HR), non-invasively measured systolic blood pressure (NBP_s), and peripheral blood oxygen saturation (SpO_2) – and their respective alarm thresholds. We selected these three vital parameters because they have the highest temporal resolution within MIMIC-III. Table 1 shows a complete list of used data items.

2.2 Data Cleaning

Vital parameters cannot assume arbitrary values but are limited to certain physiologically possible ranges. For example, SpO_2 can not exceed 100% and HRs

Table 1: Adapted from [3]: ITEMIDs retained while filtering CHARTEVENTS

ITEMID	Label
220045	HR
220046	HR Alarm - High
220047	HR Alarm - Low
220179	NBP_s
223751	NBP_s Alarm - High
223752	NBP_s Alarm - Low
220277	SpO_2
223769	SpO_2 Alarm - High
223770	SpO_2 Alarm - Low

above 350 bpm are rare and unsustainable. MIMIC-III contains some instances of unrealistically high or low values for the vital parameter measurements or the alarm thresholds. We consider these extreme values to be erroneous or to have a special but undocumented meaning. Either way, we can not interpret these values. From a contextual inquiry at an intensive care unit we learned that keyboards at the intensive care unit might have a rubber cover to stop germs from accumulating in and on the keyboard. But this causes the keyboard's keys to be sticky which impairs data entry and causes documentation errors. To deal with unrealistically high or low values, we remove all vital parameter measurements and alarm thresholds that we consider to be invalid from the data set. Table 2 lists lower and upper limits for each parameter type. We retain values within this range and discard values beyond this range. Figure 1 shows the distribution of measurements and thresholds before cleaning as boxplots. Many extreme outlier force the interquartile range to be a single line at the leftmost corner of the plot. A histogram would be completely unintelligible. Figure 2 shows the distributions of measurements and thresholds after cleaning as boxplots and histograms. This distribution is much more reasonable. The measurement histogram exhibits a bell-shaped distribution that is only slightly positively skewed.

Table 2: Adapted from [12]: Physiologically possible ranges for the vital parameters considered in this work.

Parameter	Lower Limit	Upper Limit
HR	0 bpm	350 bpm
NBP_s	0 mmHg	375 mmHg
SpO_2	0%	100%

For thresholds one condition must always hold true: The low threshold must always be lower than the high threshold. This is an invariant for all vital parameters. If this condition is not met, every measurement of the vital parameter – regardless of its value – will trigger an alarm event; which defies the purpose

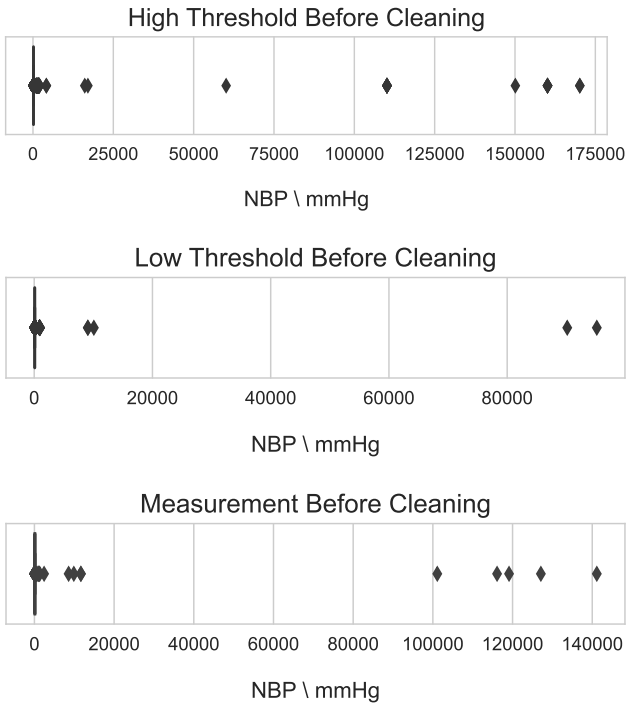


Fig. 1: Adapted from [3]: Boxplots showing the distribution of NBP_s high alarm thresholds, low alarm thresholds, and measurements before cleaning. The distribution is vastly skewed with the valid range barely visible at the far left corner and a wide range of outliers.

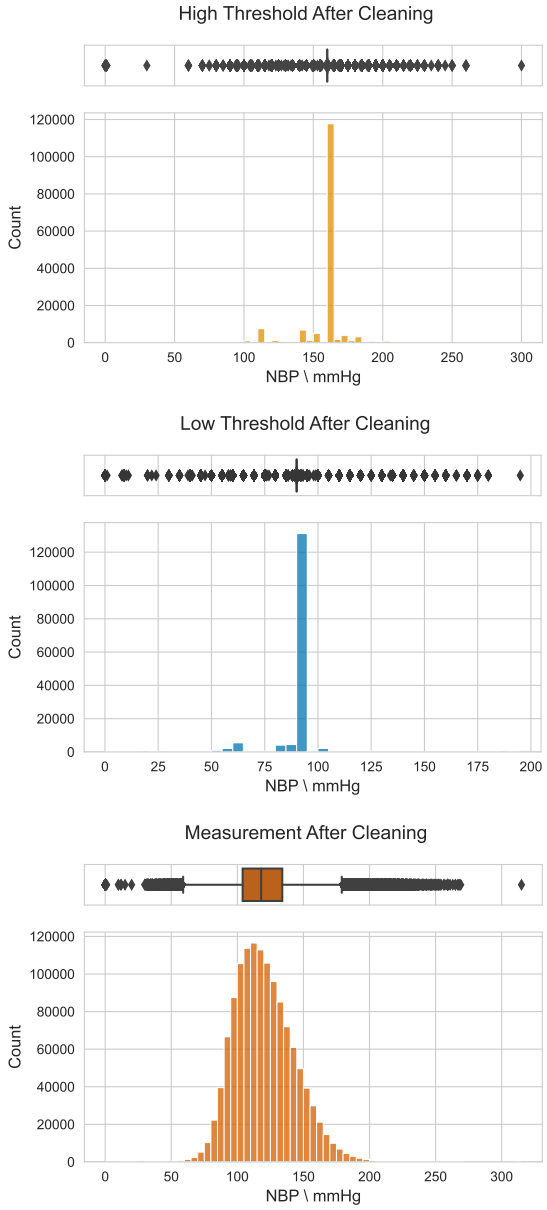


Fig. 2: Adapted from [3]: Boxplots and histograms showing the distribution of NBP_s high alarm thresholds, low alarm thresholds, and measurements after cleaning. With outliers removed, the distribution looks much more reasonable and especially the measurement values are almost normally distributed with only a slight positive skew.

of alarms altogether. MIMIC-III violates this invariant occasionally: Although high and low thresholds are always recorded simultaneously, the high threshold is sometimes lower than the low threshold. We address this problem in two ways:

Exact threshold swaps When thresholds are exactly swapped, the low threshold takes the value of the high threshold and vice versa. We correct this by swapping the set of thresholds back to normal (Figure 3).

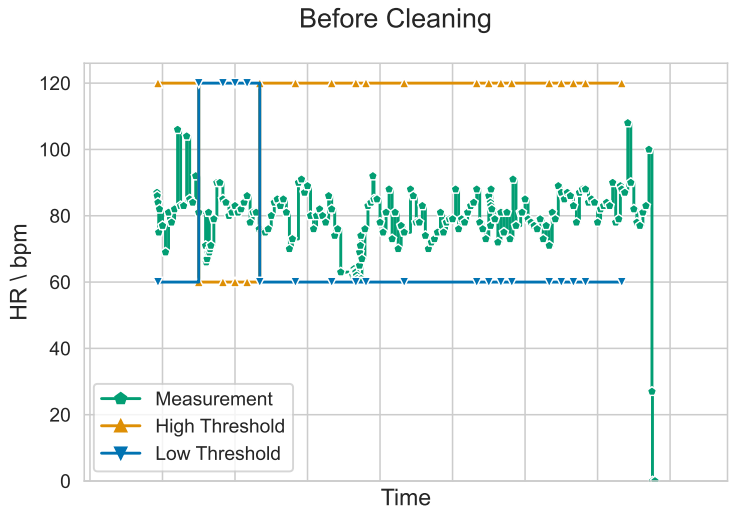
Threshold overlaps When thresholds overlap but we cannot identify an exact swap, we just remove the erroneous threshold value and carry over the previously active threshold (Figure 4).

2.3 Extracting Alarm Events

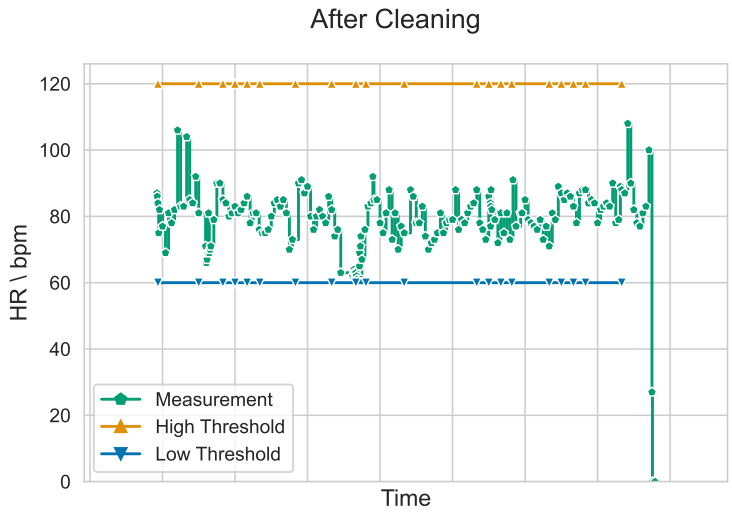
Now that we addressed all data quality issues, we can extract alarm events. We do this by using algorithm 1 on MIMIC-III's CHARTEVENTS table. First, we split the CHARTEVENTS table by ICUSTAY – a patient's single stay at the intensive care unit (one patient might be at the same intensive care unit multiple times throughout his or her life). Then, we compare each of the patient's vital parameter measurements to the high and low thresholds active at the time of measurement. If the measurement exceeds the high threshold or is below the low thresholds, the algorithm yields and alarm event. A major drawback of this methods is that the sampling frequency of the measurements influences the number of alarms: HR and SpO_2 are measured and recorded more often than NBP_s, hence more alarms are extracted by the algorithm. But this does not show that the patient actually spend more *time* with unhealthy NBP_s as compared to unhealthy HR or SpO_2.

2.4 Resampling Vital Parameters

By extracting the alarm events, we created the labels for our forecasting system. Now we still have to prepare the vital parameter time series to be used as input for the statistical and machine learning models described in section 3. The input data preparation involves two steps: resampling and chunking. Our approach is that we want to use time-series models to forecast on the patient's vital parameter measurements – as a regression task. Time-series models rely on constant sampling frequencies in the time series. But unlike eICU CRD with its vitalPeriodic table, MIMIC-III's CHARTEVENTS table does not record measurements with a constant sampling frequency but reports all charted events and data (incl. measurements) only sporadically. We first have to establish constant sampling frequencies for all vital parameter measurements so that time-series models can work with them. Whenever possible, we resample to $f_s = 1h^{-1}$ since this is approximately equal to the median sampling frequency of the measurements. To do this, we employ three different resampling methods: minimum, maximum, and median resampling. Later-on, we will compare how the different resampling methods influence the forecasting performance.



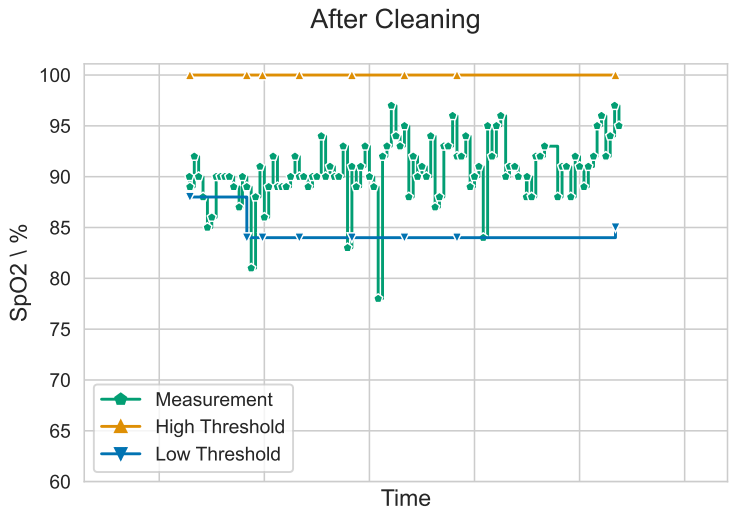
(a) Exactly swapped low and high thresholds before correction. Every measurement in the time period where the thresholds are swapped will produce an alarm.



(b) A data cleaning step removes the exact threshold swap thus rectifying the alarm thresholds. Now we will not recognise any alarm events in the respective time period.

Fig. 3: Adapted from [3]: Example for an exact threshold swap correction.

(a) In this case, the thresholds overlap without being exactly swapped. Here, the unreasonable low value for the high threshold would result in all measurements in the respective period of time triggering a high threshold alarm.



(b) Threshold overlap was corrected by removing the responsible alarm threshold settings. After correction, the measurements do not trigger any high alarms in the respective period of time.

Fig. 4: Adapted from [3]: Example for threshold overlap correction.

```
Data: MIMIC-III CHARTEVENTS
Result: List of Alarm Events
foreach ICUSTAY do
    foreach Parameter do
        msmts := measurements for Parameter and ICUSTAY;
        highs := high threshold settings for Parameter and ICUSTAY;
        lows := low threshold settings for Parameter and ICUSTAY;
        foreach high in highs do
            foreach msmt in msmts do
                if time(high) <= time(msmt) < time(high+1) then
                    if value(msmt) > value(high) then
                        Return a high alarm event at msmt;
                    end
                end
            end
        end
        foreach low in lows do
            foreach msmt in msmts do
                if time(low) <= time(msmt) < time(low+1 then
                    if value(msmt) < value(low) then
                        Return a low alarm event at msmt;
                    end
                end
            end
        end
    end
end
```

Algorithm 1: Taken from [3]: Algorithm for extracting alarm events from measurements and thresholds.

2.5 Chunking to Avoid Data Gaps

Occasionally, there are larger gaps in the vital parameter measurements. We assume that this is because the patient is not at the intensive care unit but in the operation theatre or some other ward. In these cases we do not attempt resampling but revert to chunking: We subdivide the patient's time-series data along the data gaps and treat these gaps separately as if they would belong to different ICUSTAYs for the same patient. This chunking procedure has the disadvantage that the model has to re-learn every time a data gap occurs and cannot provide alarm forecasts for some hours. We argue that our chunking methods makes sense anyway since the patient might be in a completely different state after surgery then before.

3 Alarm Forecasting

We want to forecast threshold alarms. To do this, we use time-series models to forecast the vital parameters measurements as a regression task. Then we use the forecast vital parameters to check whether they will be above the high threshold or below the low threshold in the near future. For the time-series models we compare two different model paradigms: Statistical models that do not need a separate set of training data and machine learning models that we first train on a dedicated training data set (a part of the original data set). We frame the problem as a regression task to ensure comparability between the model paradigms since the statistical models cannot perform classification right away. Also, all models are provided with the same set of features: Although we could improve machine learning models by adding more features we refrained from doing so. With all the data set issues listed in section 2, superb model performance is not the goal of this work. We rather want to provide a proof of concept and compare different model paradigms using the same data.

Experiment Setup The basic setup is the same for both model paradigms: We provide the model with either 12 or 30 timesteps (lags) of vital parameter. This is equivalent to 12 hours or 30 hours of intensive care unit stay data as input. We expect a vital parameter forecast for the 13th or 31st lag. If the vital parameter forecast is above the high threshold, a high alarm is forecast. If the vital parameter forecast is below the low threshold, a low alarm is forecast. Otherwise, no alarm is forecast. Finally, we compare the forecast with the actual alarm situation as established in subsection 2.3.

Evaluation For evaluation, we face a similar problem as Clifford et al. when they posed the 2015 PhysioNet/Computing in Cardiology Challenge which aims at reducing false arrhythmia alarms in the intensive care unit [5]. For Clifford et al., false negatives were much worse than false positives since no arrhythmia should pass unnoticed. They developed a metric that accounts for this imbalance and penalised false negatives five times more heavily than false positives (Equation 1). For us the situation is vice versa: We already noticed in section 1 that

some alarms cannot be forecast because they are the result of an acute event and not a continued trend – false negatives are to be expected. But we absolutely want to avoid increasing the workload for medical staff, hence we want to avoid false positives. We adapted the evaluation score from Clifford et al. to fit our problem (Equation 2). We also removed the true negatives from the equation since we are not interested in no-alarm situations where there is neither an alarm not a forecast for an alarm.

$$\text{Clifford's evaluation score} = \frac{TP + TN}{TP + TN + FP + 5 \cdot FN} \tag{1}$$

$$\text{our evaluation score} = \frac{TP}{TP + 5 \cdot FP + FN} \tag{2}$$

Statistical Models Statistical time-series models forecast without training on other time-series in advance and thus without prior knowledge through similar-time series. We use the autoregressive integrated moving average (ARIMA) model and the autoregressive integrated moving average with exogenous variables (ARIMAX) model. ARIMA uses only one time-series as input: the *endogenous* series. For ARIMA, we compare median resampling with either minimum resampling for low alarms or maximum resampling for high alarms. ARIMAX has another time-series – the *exogenous* series – in addition to the endogenous series. For ARIMAX, we use minimum resampling for low alarms and maximum resampling for high alarms as endogenous series. As exogenous series, we use the median-resampled vital parameter series for both high and low alarms. Additionally, we modulate the input size resulting in six different model configurations (Table 3).

Table 3: Adapted from [4]: ARIMA and ARIMAX models.

Model ID	Input Size	Model Type	Endog.
A_01_12	12	ARIMA	Median
A_02_12	12	ARIMA	Min/Max
A_03_12	12	ARIMAX	Min/Max
A_01_30	30	ARIMA	Median
A_02_30	30	ARIMA	Min/Max
A_03_30	30	ARIMAX	Min/Max

Machine Learning Models Unlike statistical model, machine learning models undergo a separate training phase before they can make predictions. We use the class of recurrent neural networks (RNNs), since these are usually used on time-series [18,19,7]. Specifically, we compare vanilla RNNs, gated recurrent units (GRUs), and long short-term memory neural networks (LSTMs). With all model

types, we use 80% of each chunk as training data and 20% as test data for assessing model performance. Otherwise, we use the same setup as for the statistical models: 12 or 30 lags as input and the 13th or 31st lag to be forecast and then checked against the alarm threshold. We then repeat this to cover the whole chunk. Additionally, we also want to test if and how scaling the input data influences the model's performance. We compare:

1. no scaling (suffix n)
2. standard scaling: $x_{scaled} = \frac{x-\mu}{\sigma}$ (suffix s1)
3. min-max scaling: $x_{scaled} = \frac{x-min}{max-min}$ (suffix s2)

Table 4 lists a machine learning model configurations.

Table 4: Adapted from [4]: Machine learning (ML) models. Standard scaling is indicated by the suffix "s1". Min-max scaling by the suffix "s2". If no scaling is performed, the suffix is "n" for "non-scaled".

Model ID	Scaling	Model Type	Endog.
LS_01_s1	Standard	LSTM	Median
LS_02_s1	Standard	LSTM	Min/Max
GR_01_s1	Standard	GRU	Median
GR_02_s1	Standard	GRU	Min/Max
RN_01_s1	Standard	RNN	Median
RN_02_s1	Standard	RNN	Min/Max
LS_01_s2	Min-Max	LSTM	Median
LS_02_s2	Min-Max	LSTM	Min/Max
GR_01_s2	Min-Max	GRU	Median
GR_02_s2	Min-Max	GRU	Min/Max
RN_01_s2	Min-Max	RNN	Median
RN_02_s2	Min-Max	RNN	Min/Max
LS_01_n	None	LSTM	Median
LS_02_n	None	LSTM	Min/Max
GR_01_n	None	GRU	Median
GR_02_n	None	GRU	Min/Max
RN_01_n	None	RNN	Median
RN_02_n	None	RNN	Min/Max

4 Results

In this section, we present the individual model performances. We first compare the statistical models among each other. Then, we compare the ML models among each other. Finally, we compare both model paradigms to each other.

Figure 5 compares how the input size influences the models performance across statistical models. As expected, longer input sequences usually yield better

model performance. This suggests that predictions will improve the longer the patient stays at the intensive care unit. To get better predictions earlier, we need to increase the sampling frequency. This is not possible with MIMIC-III but eICU CRD might be promising as long as we find a way to add alarm data.

Figure 6 compares the best statistical models among each other, contrasting high and low alarms for different vital parameters separately. We highlighted the best performing model for each alarm type in a more saturated colour. The performance varies greatly with alarm type. High alarms are generally more foreseeable, at least for HR and NBP_s. Peak performance for high and low alarms is not necessarily achieved with the same model, for example in HR and SpO_2. For future work it might be best to consider high and low alarms as completely different endpoints and not trying to build one model for multiple alarm types.

Figure 7 compares all machine learning models among each other, contrasting model type. No single model type stands out as alarm type, vital parameter, and scaling obviously influence the performance of all models. But it seems that scaling has a clearly negative effect on the models' performance. This calls for further investigation in the next figures.

Figure 8 compares the effect of different scaling methods on machine learning models. The figure confirms that scaling negatively influences the models' performance. The negative influence is most obvious in SpO_2 models. As with statistical models (Figure 6), alarm type influences the performance, but differently. For HR, low alarms are more foreseeable. For NBP_s, high alarms are more foreseeable. For SpO_2, the influence of scaling is serve and obscures differences between high and low alarms. But looking at SpO_2 models with no scaling, high alarms are more foreseeable as per this model.

Figure 9 compares all machine learning models and contrasts performances for high and low alarms. Again, we highlighted the best performing models for each alarm type in a more saturated colour. Multiple models exhibit the same peak performance for SpO_2 high alarms. Otherwise, this figure clearly shows that GRU models with median resampling show an overall superior performance. This is an important direction for future research.

Finally, Figure 10 compares the best performing statistical models with the best performing machine learning models. Mostly, machine learning models outperform statistical models, especially in the SpO_2 use case. Figure 11 shows why this is the case: The confusion matrix reveals that machine learning models are much better at avoiding false positives. Since false positives are penalised five times more heavily as per our evaluation metrics, this is a major advantage for the model in this specific scenario. Machine learning models – with their prior knowledge through training – do not tend to forecast extreme values as much as statistical model that do not have prior knowledge on the domain. As extreme vital parameter forecasts cause alarms forecasts, this prior knowledge helps avoiding false positive alarms and improves evaluation scores.

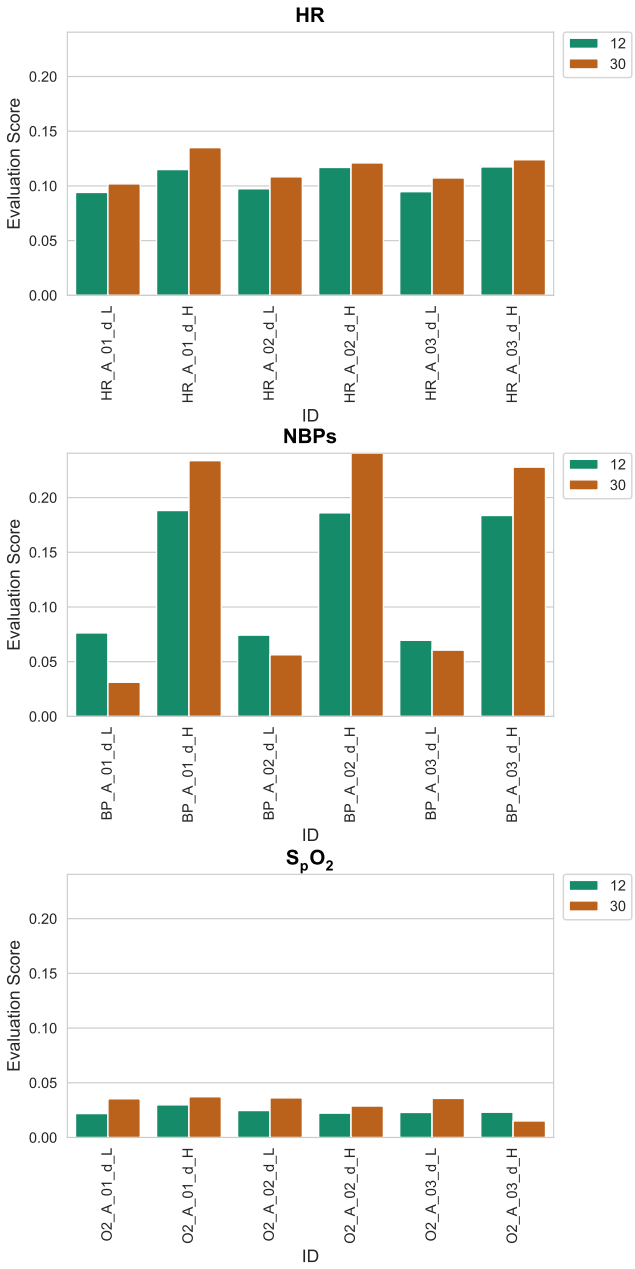


Fig. 5: Comparison of train sizes for statistical models (ARIMA and ARIMAX). For all parameters and model we compare a train size of 12 lags with a train size of 30 lags both for high alarms (suffix _H) and low alarms (suffix _L).

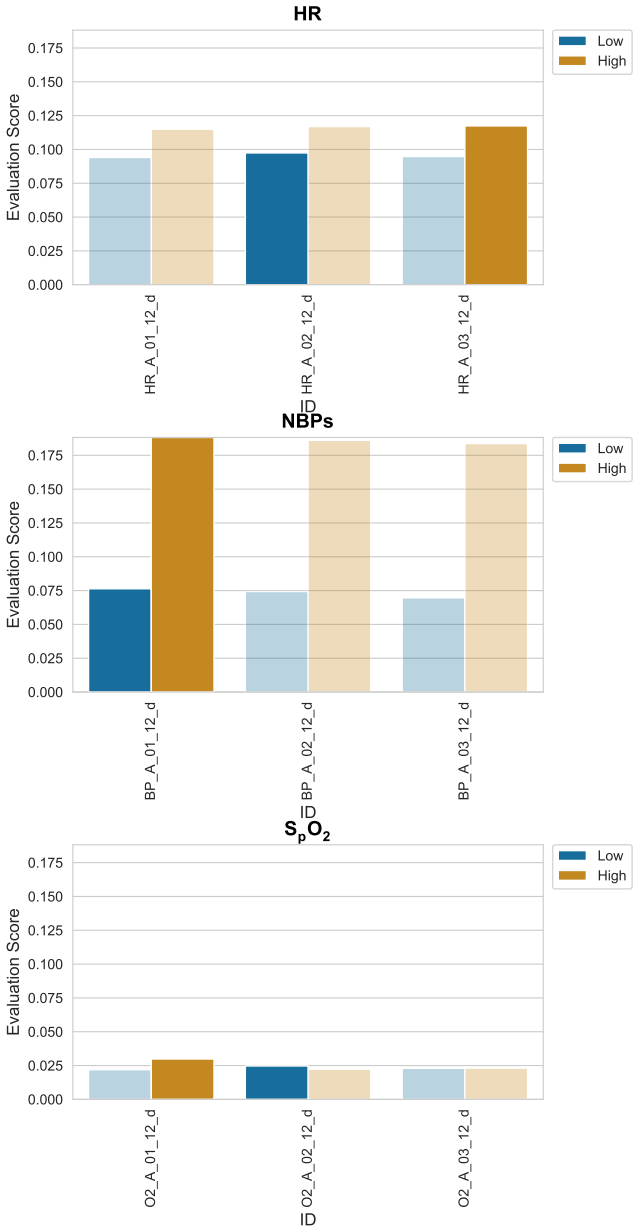


Fig. 6: Comparison of alarm types (high alarm and low alarm) for statistical models across all vital parameters.

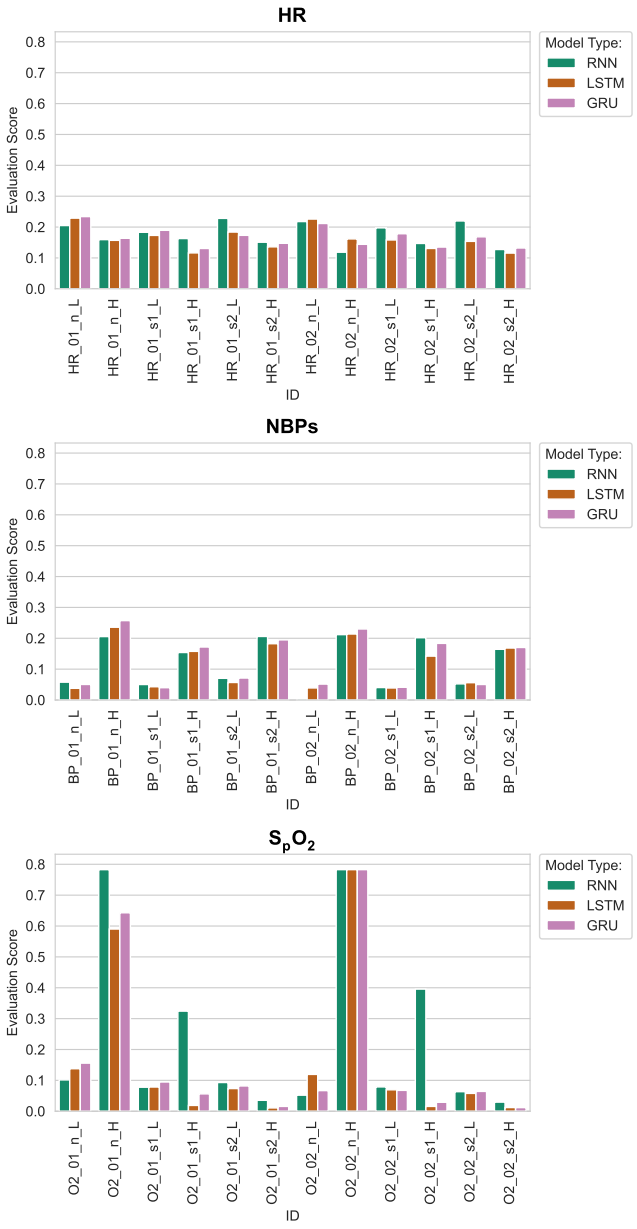


Fig. 7: Comparison of ML model types vanilla RNN, LSTM and GRU with different configurations across vital parameters and alarm types (suffix _H for high alarms and suffix _L for low alarms).

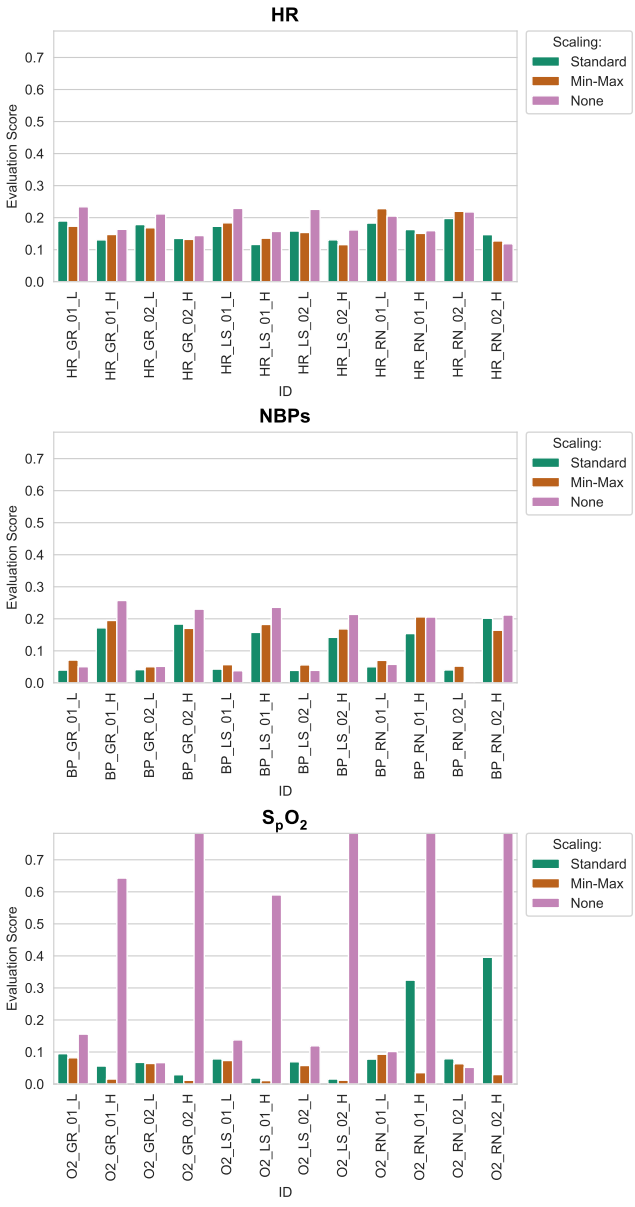


Fig. 8: Comparison of ML models executed with different scaling methods applied (Standard, Min-Max) or without scaling (None) across vital parameters and alarm types (suffix _H for high alarms and suffix _L for low alarms).

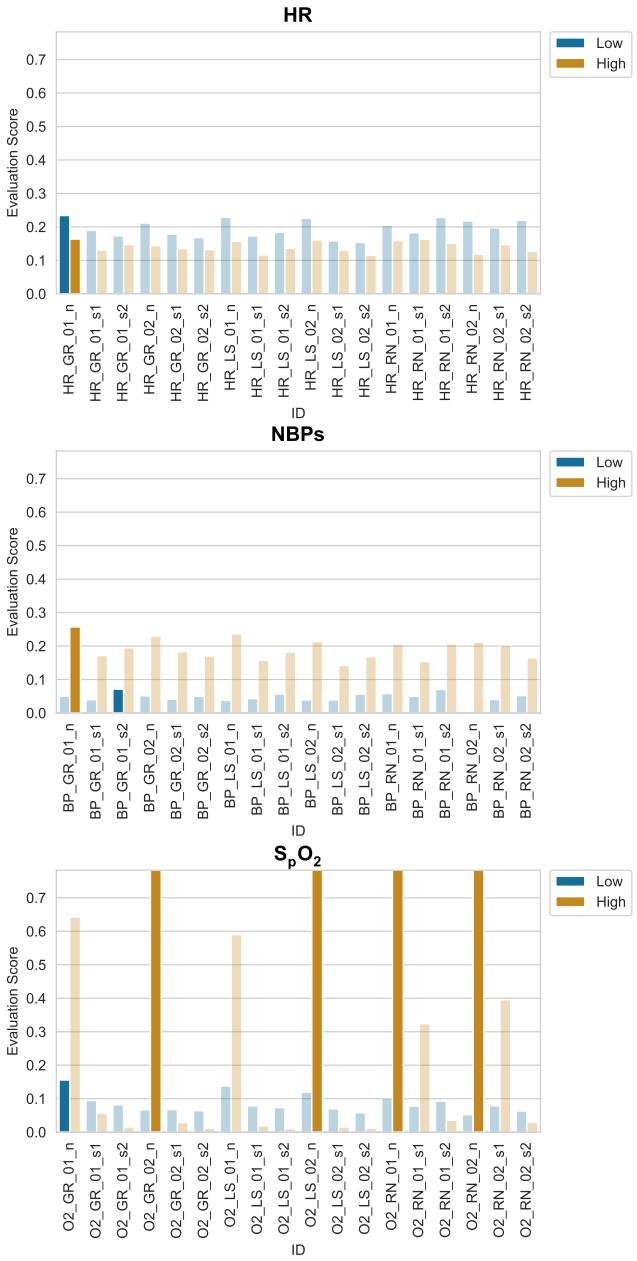


Fig. 9: Selection of best ML models. Models with model type GRU and median resampled chunks as endogenous input variable always perform best (except for high alarm forecasting of SpO_2).

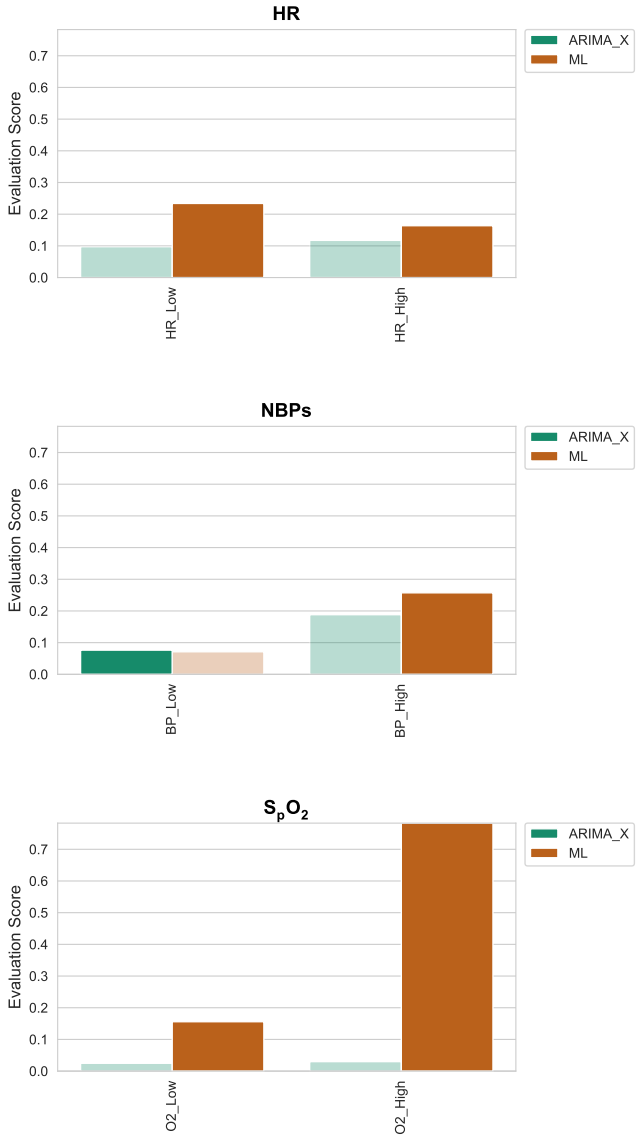


Fig. 10: Comparison of best performing statistical models to best performing ML models.

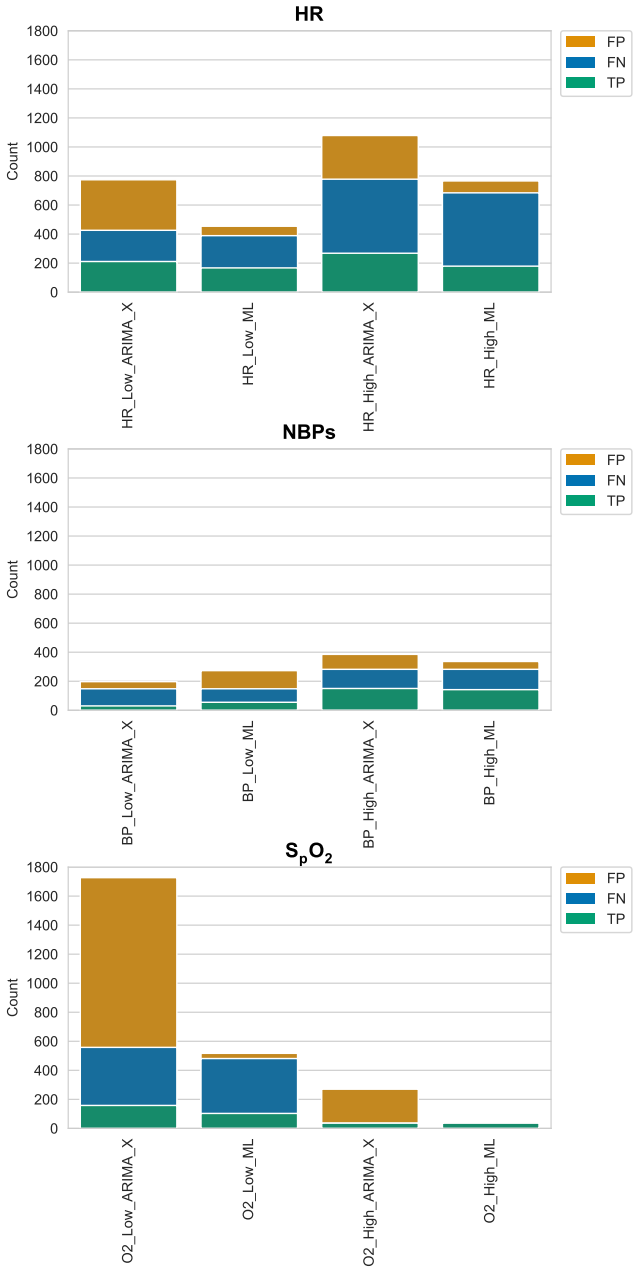


Fig. 11: Comparison of confusion matrix values (false positives, false negatives, and true positives; not showing true negatives) of best performing statistical models to best performing ML models.

5 Discussion

We have shown a method to forecast threshold alarms. This, however, is limited to the share of alarms that can be forecast because of a continued trend. Acute events – for example sudden onset of cardiac arrhythmia – cannot be foreseen by our method. With our method, we can transform a share of the alarms into scheduled tasks. Thus removing the urgency from the situation and reducing the alarm load. Also, forecasting alarms buys staff more time to treat critical conditions which also benefits patients. This way, patients can get treatment even before their conditions becomes overly critical.

Alarm fatigue is a well-known problem in medicine with many detrimental effects on patients and staff. From existing research on alarm fatigue, it is perfectly clear that we must reduce the number of alarms. But we do not know yet how we can reduce the number of alarms without risking to overlook a critical condition and sacrificing patient safety. Paine et al. systematically reviewed existing literature on alarm fatigue and compiled a list of alternative approach to reduce the alarm load: Widening alarm thresholds and introducing alarm delays can reduce the total number of alarms but might have adverse safety outcomes. Using disposable electrocardiographic lead wires and changing electrodes daily will reduce measurement errors and reduce technically false alarms without endangering patients but increases staff workload and monitoring costs. Finally, for some interventions the safety outcomes are yet unclear, for example changing alarm sounds and presentation, personal alarming through pagers or mobile phones, and focusing monitoring on high-risk patients while relaxing monitoring on low-risk patients. Our forecasting approach does not endanger patients, produces no additional cost, and – by emphasising low false positive rates – does not increase staff workload.

The method we proposed in this work is only a proof of concept. We showed that the approach works in principle and we found important indications for future work. The methods, however, is not yet ready for productive use. The most striking limitation is a data set issue. As we already mentioned in section 2, MIMIC-III is the only clinical data set that contains alarm data. But its low temporal resolution regarding vital parameter measurements vastly limits MIMIC-III's usability. With unsteady sampling frequencies of $f_s \approx 1\text{h}^{-1}$, forecasting is difficult and limited. Other clinical data sets feature higher and steady sampling frequencies – for example eICU CRD with steady $f_s \approx 12\text{h}^{-1}$ – but lack alarm data altogether. For this work, we chose MIMIC-III and tried to cope with the low temporal resolution. To circumvent the low sampling frequency issue, future work can focus on vital parameter forecasting using eICU CRD, omitting the actual alarm event forecasting or using simulated alarm thresholds. Another approach for future work could be framing the problem as a classification task rather than a regression task. Through this work, we already know that machine learning models outperform statistical models. A possible next step can be to remove the indirection of forecasting the vital parameter measurement and forecast the alarm right away.

Reducing the number of false alarms at the intensive care unit and counteracting alarm fatigue is very much necessary according to domain experts. The models we proposed are far from providing perfect alarm forecasts. This is mostly due to the data set issues described above. But there is huge potential for this approach to alleviate alarm fatigue in the future with better data sets featuring high temporal vital parameter resolution and precise alarm data.

Acronyms

AmsterdamUMCdb Amsterdam University Medical Centers Database
ARIMA autoregressive integrated moving average
ARIMAX autoregressive integrated moving average with exogenous variables
eICU CRD eICU Collaborative Research Database
GRU gated recurrent unit
HiRID High Time Resolution ICU Data Set
HR heart rate
LSTM long short-term memory neural network
MIMIC-III 3rd version of the Medical Information Mart for Intensive Care
ML machine learning
NBP_s non-invasively measured systolic blood pressure
RNN recurrent neural network
SpO_2 peripheral blood oxygen saturation

References

1. Baker, C.F., Garvin, B.J., Kennedy, C.W., Polivka, B.J.: The effect of environmental sound and communication on CCU patients' heart rate and blood pressure. Research in Nursing & Health **16**(6), 415–421 (Dec 1993). https://doi.org/10.1002/nur.4770160605
2. Bennun, I.: Intensive care unit syndrome: A consideration of psychological interventions. British Journal of Medical Psychology **74**(3), 369–377 (Sep 2001). https://doi.org/10.1348/000711201161046
3. Chromik, J., Pfitzner, B., Ihde, N., Michaelis, M., Schmidt, D., Klopfenstein, S., Poncette, A.S., Balzer, F., Arnrich, B.: Extracting Alarm Events from the MIMIC-III Clinical Database. In: 15th International Conference on Health Informatics. pp. 328–335 (2022). https://doi.org/10.5220/0010767200003123
4. Chromik, J., Pfitzner, B., Ihde, N., Michaelis, M., Schmidt, D., Klopfenstein, S., Poncette, A.S., Balzer, F., Arnrich, B.: Forecasting Thresholds Alarms in Medical Patient Monitors using Time Series Models. In: 15th International Conference on Health Informatics. pp. 26–34 (2022). https://doi.org/10.5220/0010767300003123
5. Clifford, G.D., Silva, I., Moody, B., Li, Q., Kella, D., Shahin, A., Kooistra, T., Perry, D., Mark, R.G.: The PhysioNet/Computing in Cardiology Challenge 2015: Reducing false arrhythmia alarms in the ICU. In: 2015 Computing in Cardiology Conference (CinC). pp. 273–276. IEEE, Nice, France (Sep 2015). https://doi.org/10.1109/CIC.2015.7408639

6. Cvach, M.: Monitor Alarm Fatigue: An Integrative Review. Biomedical Instrumentation & Technology **46**(4), 268–277 (Jul 2012). https://doi.org/10.2345/0899-8205-46.4.268
7. Dai, X., Liu, J., Li, Y.: A recurrent neural network using historical data to predict time series indoor PM2.5 concentrations for residential buildings. Indoor Air **31**(4), 1228–1237 (2021). https://doi.org/10.1111/ina.12794
8. Drew, B.J., Harris, P., Zègre-Hemsey, J.K., Mammone, T., Schindler, D., Salas-Boni, R., Bai, Y., Tinoco, A., Ding, Q., Hu, X.: Insights into the Problem of Alarm Fatigue with Physiologic Monitor Devices: A Comprehensive Observational Study of Consecutive Intensive Care Unit Patients. PLoS ONE **9**(10), e110274 (Oct 2014). https://doi.org/10.1371/journal.pone.0110274
9. Fife, D., Rappaport, E.: Noise and hospital stay. American Journal of Public Health **66**(7), 680–681 (Jul 1976). https://doi.org/10.2105/ajph.66.7.680
10. Granberg, A., Bergbom Engberg, I., Lundberg, D.: Patients' experience of being critically ill or severely injured and cared for in an intensive care unit in relation to the ICU syndrome. Part I. Intensive & Critical Care Nursing **14**(6), 294–307 (Dec 1998). https://doi.org/10.1016/s0964-3397(98)80691-5
11. Hagerman, I., Rasmanis, G., Blomkvist, V., Ulrich, R., Eriksen, C.A., Theorell, T.: Influence of intensive coronary care acoustics on the quality of care and physiological state of patients. International Journal of Cardiology **98**(2), 267–270 (Feb 2005). https://doi.org/10.1016/j.ijcard.2003.11.006
12. Harutyunyan, H., Khachatrian, H., Kale, D.C., Ver Steeg, G., Galstyan, A.: Multitask learning and benchmarking with clinical time series data. Scientific Data **6**(1), 96 (Jun 2019). https://doi.org/10.1038/s41597-019-0103-9
13. Hyland, S.L., Faltys, M., Hüser, M., Lyu, X., Gumbsch, T., Esteban, C., Bock, C., Horn, M., Moor, M., Rieck, B., Zimmermann, M., Bodenham, D., Borgwardt, K., Rätsch, G., Merz, T.M.: Early prediction of circulatory failure in the intensive care unit using machine learning. Nature Medicine **26**(3), 364–373 (Mar 2020). https://doi.org/10.1038/s41591-020-0789-4
14. Johnson, A.E.W., Pollard, T.J., Shen, L., Lehman, L.W.H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L.A., Mark, R.G.: MIMIC-III, a freely accessible critical care database. Scientific Data **3**, 160035 (May 2016). https://doi.org/10.1038/sdata.2016.35
15. Minckley, B.B.: A study of noise and its relationship to patient discomfort in the recovery room. Nursing Research **17**(3), 247–250 (1968 May-Jun)
16. Morrison, W.E., Haas, E.C., Shaffner, D.H., Garrett, E.S., Fackler, J.C.: Noise, stress, and annoyance in a pediatric intensive care unit. Critical Care Medicine **31**(1), 113–119 (Jan 2003). https://doi.org/10.1097/00003246-200301000-00018
17. Murthy, V.S., Malhotra, S.K., Bala, I., Raghunathan, M.: Detrimental effects of noise on anaesthetists. Canadian Journal of Anaesthesia = Journal Canadien D'anesthesie **42**(7), 608–611 (Jul 1995). https://doi.org/10.1007/BF03011878
18. Mussumeci, E., Codeço Coelho, F.: Large-scale multivariate forecasting models for Dengue - LSTM versus random forest regression. Spatial and Spatio-Temporal Epidemiology **35**, 100372 (Nov 2020). https://doi.org/10.1016/j.sste.2020.100372
19. Pathan, R.K., Biswas, M., Khandaker, M.U.: Time series prediction of COVID-19 by mutation rate analysis using recurrent neural network-based LSTM model. Chaos, Solitons, and Fractals **138**, 110018 (Sep 2020). https://doi.org/10.1016/j.chaos.2020.110018
20. Pisani, M.A., Friese, R.S., Gehlbach, B.K., Schwab, R.J., Weinhouse, G.L., Jones, S.F.: Sleep in the intensive care unit. American Journal of Respiratory and Critical

Care Medicine **191**(7), 731–738 (Apr 2015). https://doi.org/10.1164/rccm.201411-2099CI
21. Pollard, T.J., Johnson, A.E.W., Raffa, J.D., Celi, L.A., Mark, R.G., Badawi, O.: The eICU Collaborative Research Database, a freely available multi-center database for critical care research. Scientific Data **5**, 180178 (Sep 2018). https://doi.org/10.1038/sdata.2018.178
22. Poncette, A.S., Mosch, L., Spies, C., Schmieding, M., Schiefenhövel, F., Krampe, H., Balzer, F.: Improvements in Patient Monitoring in the Intensive Care Unit: Survey Study. Journal of Medical Internet Research **22**(6), e19091 (Jun 2020). https://doi.org/10.2196/19091
23. Ryherd, E.E., Waye, K.P., Ljungkvist, L.: Characterizing noise and perceived work environment in a neurological intensive care unit. The Journal of the Acoustical Society of America **123**(2), 747–756 (Feb 2008). https://doi.org/10.1121/1.2822661
24. Schmid, F., Goepfert, M.S., Kuhnt, D., Eichhorn, V., Diedrichs, S., Reichenspurner, H., Goetz, A.E., Reuter, D.A.: The Wolf Is Crying in the Operating Room: Patient Monitor and Anesthesia Workstation Alarming Patterns During Cardiac Surgery. Anesthesia & Analgesia **112**(1), 78–83 (Jan 2011). https://doi.org/10.1213/ANE.0b013e3181fcc504
25. Thoral, P.J., Peppink, J.M., Driessen, R.H., Sijbrands, E.J.G., Kompanje, E.J.O., Kaplan, L., Bailey, H., Kesecioglu, J., Cecconi, M., Churpek, M., Clermont, G., van der Schaar, M., Ercole, A., Girbes, A.R.J., Elbers, P.W.G.: Sharing ICU Patient Data Responsibly Under the Society of Critical Care Medicine/European Society of Intensive Care Medicine Joint Data Science Collaboration: The Amsterdam University Medical Centers Database (AmsterdamUMCdb) Example*. Critical Care Medicine **49**(6), e563–e577 (Jun 2021). https://doi.org/10.1097/CCM.0000000000004916
26. Topf, M., Dillon, E.: Noise-induced stress as a predictor of burnout in critical care nurses. Heart & Lung: The Journal of Critical Care **17**(5), 567–574 (Sep 1988)
27. Wilken, M., Hüske-Kraus, D., Klausen, A., Koch, C., Schlauch, W., Röhrig, R.: Alarm Fatigue: Causes and Effects. Studies in Health Technology and Informatics **243**, 107–111 (2017)
28. Wysocki, A.B.: The effect of intermittent noise exposure on wound healing. Advances in Wound Care: The Journal for Prevention and Healing **9**(1), 35–39 (1996 Jan-Feb)