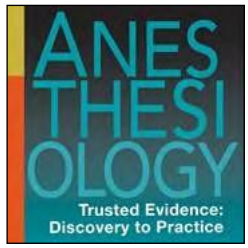
ANESTHESIOLOGY
Trusted Evidence: Discovery to Practice®

Journal Pre-proof



A Comparison of 5 Algorithmic Methods and Machine Learning Pattern Recognition for Artifact Detection in Electronic Records of 5 Different Vital Signs: A Retrospective Analysis

Mathias Maleczek, MD, Daniel Laxar, MD, Lorenz Kapral, M.Sc, Melanie Kuhrn, Predoc, Yannic Abulez, Christoph Dibiasi, MD, Oliver Kimberger, MD, M.Sc., MBA.

DOI: https://doi.org/10.1097/ALN.0000000000004971

To appear in: Anesthesiology

Submitted for publication: January 19, 2023

Accepted for publication: February 22, 2024

Please cite this article as: Maleczek M, Laxar D, Kapral L, Kuhrn M, Abulez Y, Dibiasi C, Kimberger O: A Comparison of 5 Algorithmic Methods and Machine Learning Pattern Recognition for Artifact Detection in Electronic Records of 5 Different Vital Signs: A Retrospective Analysis. Anesthesiology. 2024; https://doi.org/10.1097/ALN.0000000000004971

Anesthesiology Publish Ahead of Print

DOI: 10.1097/ALN.0000000000004971

A Comparison of 5 Algorithmic Methods and Machine Learning Pattern Recognition for Artifact Detection in Electronic Records of 5 Different Vital Signs: A Retrospective Analysis

- Mathias Maleczek, MD[a,b*], Researcher, mathias.maleczek@meduniwien.ac.at
- Daniel Laxar Daniel, MD[a,b], Researcher, daniel.laxar@dhps.lbg.ac.at
- Lorenz Kapral, M.Sc[b], Researcher, Lorenz.kapral@dhps.lbg.ac.at
- Melanie Kuhrn, Predoc[b]
- Yannic Abulez, Researcher[b]
- Christoph Dibiasi, MD[a,b], Researcher, christoph.dibiasi@meduniwien.ac.at
- Oliver Kimberger[a,b,c], MD, M.Sc., MBA, Professor, oliver.kimberger@meduniwien.ac.at

[a]Department of Anesthesiology, Intensive Care Medicine and Pain Medicine, Medical University of Vienna, Vienna, Austria

[b]Ludwig Boltzmann Institute for Digital Health and Patient Safety, Medical University of Vienna, Vienna, Austria

Corresponding author: Mathias Maleczek, MD Department of Anesthesiology, Intensive Care Medicine and Pain Medicine, Medical University of Vienna Währinger Gürtel 18-20, 1090 Vienna, Austria mathias.maleczek@meduniwien.ac.at

Clinical trial number: Not applicable, EC #2179/2020

Previous presentations: Part of this work was presented as poster at the Euroanaesthesia Congress (17–19.12.2021 Munich, Germany).

Acknowledgments: Not applicable

Word and element counts

- Abstract: 300
- Introduction: 409
- Discussion: 1,409
- Number of Figures: 1
- Number of Tables: 5
- Number of Appendices: 0
- Number of Supplemental Digital Files: 5

Abbreviated title: Comparison of Artifact Detection Methods

Summary statement: Not applicable

Funding statement: This study was funded only by departmental funds. The article processing charge was funded by a University Licensing Agreement through the Medical University of Vienna.

Conflict of interest: The authors declare no competing interests.

Code availability: Code of the used algorithms are found in Supplement 2&3; Full code is available from the corresponding author upon reasonable request.

- Abstract

  - Background

Research on electronic health record physiological data is common, invariably including artifacts. Traditionally, these artifacts have been handled using simple filter techniques. The authors hypothesized different artifact detection algorithms, including machine learning, may be necessary to provide optimal performance for various vital signs and clinical contexts.

  - Materials and Methods

In a retrospective single center study, intraoperative OR and ICU electronic health record datasets including heart rate, oxygen saturation, blood pressure, temperature, and capnometry were included. All records were screened for artifacts by at least two human experts. Classical artifact detection methods (cutoff, multiples of standard deviation (z-value), interquartile range, and local outlier factor) and a supervised learning model implementing long short-term memory neural networks were tested for each vital sign against the human expert reference dataset. For each artifact detection algorithm, sensitivity and specificity were calculated.

  - Results

A total of 106 (53 operating room and 53 ICU) patients were randomly selected, resulting in 392,808 data points. Human experts annotated 5,167 (1.3%) data points as artifacts. The artifact detection algorithms demonstrated large variations in performance. The specificity was above 90% for all detection methods and all vital signs. The neural network showed significantly higher sensitivities than the classic methods for: heart rate (ICU: 33.6%, 95% CI: 33.1–44.6), systolic invasive blood pressure (both in the OR (62.2%, 95% CI: 57.5–71.9) and ICU (60.7%, 95% CI: 57.3–71.8), and temperature in the OR (76.1%, 95% CI: 63.6–89.7). The confidence intervals for specificity overlapped for all methods. Generally, sensitivity was low, with only the z-value for oxygen saturation in the operating room reaching 88.9%. All other sensitivities were less than 80%.

- Conclusion

No single artifact detection method consistently performed well across different vital signs and clinical settings. Neural networks may be a promising artifact detection method for specific vital signs.

- Introduction

The collection of physiological data is common in anesthesia and intensive care, and a large variety of high-resolution vital signs are generated and stored routinely in hospitals with electronic health record (EHR) systems. The availability of large datasets of vital signs offers a tremendous opportunity to conduct clinical research using data from thousands of patients. For example, large datasets of blood pressure values have enabled researchers to determine associations between intraoperative hypotension and clinically relevant outcomes, such as acute kidney injury, stroke, and myocardial injury. [1–7]

This large-scale collection of vital signs invariably includes collecting artifacts as well, with a value outside the normal ranges of the vital sign in question being likelier to be an artifact than a value with normal ranges, according to some reports.[8,9] These artifacts can originate from different factors, such as electrocautery, disconnected arterial lines, or movement of the lines.[10] Although artifacts are typically recognized easily and ignored by the treating staff in real-time, retrospective analysis of large datasets does not have documentation of this thought process. Artifact data can alter clinical outcome classification and impact descriptive and inferential analyses.[8,11,12] Artifact filtering can have a substantial impact on hypotension prevalence and a small effect on the reported association between hypotension and myocardial injury.[12] To date, most large retrospective studies use simple filter techniques, whereas some studies do not comment on artifact handling at all.[1] The most commonly used filters are cutoff filters and moving mean/median. Moving mean/median is different from all other filters, as it modifies nearly every data point.[13] It is unclear which artifact detection algorithm may be suitable for perioperative and critical care data, since large-scale studies about artifact detection in these data are rare.[8,9,13,14]

To compare different artifact filtering methods, we applied the currently used methods and augmented them with algorithms known in data science as well as a neural network specially trained for artifact filtering. [13,15] Supervised learning algorithm frameworks that implement

neural networks in pattern recognition have shown numerous successes. Since artifacts are also a type of pattern, or rather a break of a certain type of pattern, testing neural networks for artifact recognition is an obvious choice.[16,17] The main objective of this project was to provide guidance on which algorithm is best suited for filtering the artifacts of each of the most common vital parameters: heart rate, blood pressure, temperature, capnometry, and peripheral oxygen saturation. We hypothesized that a specially trained neural network would outperform classic algorithms.

- Materials and Methods
  - Study Design

The University of Vienna's ethical committee approved this study and waived the need for informed consent (reference number 2179/2020). The Medical University of Vienna is a tertiary care hospital with approximately 50,000 surgical procedures and 7,000–8,000 ICU admissions per year. We conducted a retrospective study using the Medical University of Vienna's perioperative database. The study population for this study consisted of all patients who underwent surgery between 1/1/2019 and 9/1/2020. To ensure that only complete datasets were included, ICU patients had to have at least 120 hours of records of all five vital parameters (heart rate, blood pressure, temperature, capnometry, and peripheral oxygen saturation). Device data sources for each of these parameters are described below. Surgical patients had to have at least 30 minutes of records of the five parameters.

The sample size was defined using a pragmatic approach to how much data could be annotated in a reasonable time (approximately 4.5 hours per expert reviewer participant) while providing the team with enough data points to split data and use all filtering algorithms. The sample included 53 patients admitted to the ICU and 53 patients undergoing surgery. The patients were randomly selected from a list of all included patients using a random number generator in Python.[18] To limit the amount of ICU patient data, only 120 hours of records were used per patient. Data from surgical cases were used in total.

The study followed the Enhancing the Quality and Transparency of Health Research (EQUATOR) Standards for Reporting Diagnostic Accuracy Studies (STARD) Guideline.[19] The checklist used can be found in the supplement (https://links.lww.com/ALN/D489).

- Data Sources

Data were collected from the perioperative database. The perioperative database is constantly synchronized with the Philips IntelliSpace Critical Care and Anesthesia (Philips, Amsterdam, Netherlands) electronic health record (EHR), recording all patients perioperatively and in the ICU. The database contains data on vital parameters and manually entered observations/actions by all healthcare professionals. In the operating room, discrete vital parameters are stored every 15 seconds; in the ICU, the temporal resolution is 15 minutes. No artifact recognition method is applied before saving the data; therefore, the raw parameters are saved and can be used for scientific applications. Heart rate, blood pressure, temperature, and pulse oximeter values were collected via a Draeger Infinity monitoring system consisting of both the Draeger Infinity Delta and Infinity M540 systems (Draeger, Luebeck, Germany). Capnometry values were collected via an anesthesia machine (Draeger Primus or Draeger Perseus) in the operating room. In the ICU, CO_2 is measured using Draeger monitors. The heart rate parameter studied in this analysis was the ECG heart rate; pulse rate from the arterial line was not available. During artifact filtering, the human experts had the opportunity to see the pulse rate from the oxygen saturation. Blood pressure was collected both noninvasively using a cuff and invasively using arterial lines (mainly radial). For both blood pressure signals, no further signal processing was undertaken to provide all methods with the "raw" data available in the EHR. For cases in which both invasive and noninvasive blood pressure signals were available, both were annotated by the reviewers.

- Data Processing

All available values of the five vital signs of interest were extracted from the database for the randomly selected patients. No further processing of the data was performed, except for deleting all capnometry values below 2 mmHg. These values are transferred to the EHR by the anesthesia machines as soon as they are switched on by default. Beyond that, no further changes were made to the data; no interpolation was performed.

In the first step, vital sign artifacts were annotated independently by five human experts: after their 4-month anesthesia internship in the operating theater, final-year medical students received four hours of training both as a group and individually, as well as feedback on demand. Further details can be found in the supplemental material (https://links.lww.com/ALN/D493). A web-based front-end interface was used to review the patient charts and annotate the artifacts. In this self-developed front end, all reported vital signs plus pulse rate from pulse oximetry could be seen singly or combined. Artifacts can be annotated either by clicking on single data points or by circling them to mark more than one. Across two training sessions, the experts were instructed to annotate every data point that they believed to be an artifact. This included a discussion of the most important causes: disconnected/displaced lines, blood sampling, electrocautery, patient movement, etc. Four of the experts annotated 53 patients each (equally distributed between the ICU and operating room). Each patient was annotated independently by two experts. If the two experts had conflicting annotations, the fifth member of the expert panel made the final decision regarding whether the data point was an artifact by majority vote.

- Artifact Detection Algorithms

Parallel and independent from the human artifact filtering used as the reference standard during algorithm comparison, the following artifact detection algorithms were applied to the data (invasive and non-invasive blood pressure signals had the same handling throughout):

- **Cutoff:** For the cutoff algorithm, the following ranges were defined as valid (and physiologically possible), and all values beyond these ranges were defined as artifacts: systolic blood pressure, 20–300 mmHg; mean blood pressure, 10–250 mmHg; diastolic blood pressure, 5–225 mmHg; capnometry, 5–150 mmHg; temperature, 25–45°C; heart rate, 5–300/min; and SpO_2, 0–100%.[7]
- **Z-value:** The z-value was calculated for each vital sign and for each patient. All values lying outside three multiples of the standard deviation were defined as artifacts. By using 3 standard deviations as the threshold, all values lying outside 99.73% of the mean were marked as artifacts.
- **Interquartile range:** The interquartile range was calculated for each vital sign and for each patient. All values lying outside the three multiples of the interquartile range were defined as artifacts. By using 3 interquartile ranges as the threshold, all values lying outside 99.73% of the median were marked as artifacts.
- **Local outlier factor:** The local outlier factor was calculated as described by Breuning et al. [20] Seconds for Euclidean distances were on the x-axis, and the vital sign-specific values (mmHg for blood pressure, % for SpO_2, etc.) were on the y-axis. A k = 7 was chosen primarily to identify extreme changes in the time series. A local outlier factor greater than or equal to 1.5 was used to label the data points as artifacts. [20]

Additionally, a **long short-term memory neural network** was trained using the human reference standard to show its ability to predict this standard in another part of the dataset. The dataset was transformed into batches of time series. The steps for this method included (1) normalizing the input features to an interval of [-1,1], (2) calculating the first and second derivatives of the time-dependent input values, and (3) creating a time series for each data feature within a defined time window. Any data points outside the observed period were set to 0. The dataset was then randomly split into a training set (80%) and a test set (20%) while ensuring that data from a single patient was not split. The network architecture comprised an

input layer, a long short-term memory neural network layer, and an output layer optimized for the size of the time window and batches. The process was halted once the accuracy, specificity, sensitivity, and area under the receiver operator characteristics curve (ROC) did not improve in the test set. The number of neurons was estimated based on the input size, output size, batch size, and size of the observed time interval. The loss function was applied using an entropy gradient function with the ADAM[21] optimizer. To implement these algorithms, Python 3.8[18] (primarily with the pandas[22], numpy[23], scikit-learn[24], and scipy[25] packages) was used; the code used can be found in the supplementary information (https://links.lww.com/ALN/D490, https://links.lww.com/ALN/D491).

- Statistical Analysis

The main objective of this study was to compare the sensitivity and specificity of all used artifact filtering methods to provide a guide on which algorithm is best suited for further research. For calculation of sensitivity and specificity, the human reviewer standard defined the "true values" of the presence or absence of artifacts. Sensitivity was defined as the ratio of artifacts correctly annotated by each algorithm compared to the human reviewer reference standard. Specificity was defined as the ratio of data points correctly annotated as not being an artifact consistent with the human reviewer reference standard. After all the artifact detection algorithms were applied, the results were compared to the human reference standard. For each artifact detection algorithm, true positive, true negative, false positive, and false negatives were calculated. Specificity, sensitivity, positive predictive value, and negative predictive value including 95% confidence intervals (CI) calculated using Wilson's method[26,27] were displayed per vital sign parameter and artifact detection algorithm. Comparisons of confidence intervals were done using the complete non-overlap method.[28] Observations were viewed as independent—no within-person clustering of performance was conducted. Descriptive statistics were calculated using mean and standard deviation or median and 25% and 75% quartiles, respectively, as appropriate. A formal comparison of

artifact detection methods was conducted by comparing the 95% confidence intervals of sensitivity/specificity. The null hypothesis was that there was no difference in sensitivity and specificity between the neural network and any other method. All statistical analysis was done using Python 3.8[18] as described above.

Sensitivity Analysis

As most of the tested algorithms relied on the definition of thresholds or factors, a sensitivity analysis was conducted using different thresholds. For all algorithms, receiver-operator-characteristic curves are shown, and the area under the curve was calculated. Furthermore, all key statistical figures (sensitivity, specificity, positive predictive value, and negative predictive value) are shown in the main analysis.

Defining thresholds for the cutoff method was challenging. Multiple ways of calculating a reference range are described in the literature, with a 95% CI being mostly used, especially for laboratory values.[29–31] Indeed, the literature about reference ranges of vital signs is sparse, often relying on cohort studies focusing on outcomes.[32] In oxygen saturation, for example, normal values and values not requiring treatment differ in certain patient groups (e.g., acute respiratory distress syndrome or acute myocardial infarction). [33–35]

Therefore, four additional thresholds were defined: (1) using a 95% confidence interval from the complete dataset, (2) values outside of physiological ranges, (3) values that would worry the treating healthcare professionals, and (4) values needing urgent treatment. Details can be found in Table 5.

For the z-value and interquartile range, in addition to using three as factors, all values between 2 and 3.5 were tested in 0.5 steps. For the calculation of the receiver-operator-characteristics curve, all values between 0.5 and 5 were used.

- Results

The study population consisted of 28,388 operating room patients and 1,262 ICU patients with available data for all 5 vital signs. A total of 106 patients (53 ICU and 53 operating rooms) were randomly selected from the study population. Demographic details can be found in Table 1.

For the operating room data and ICU data, the mean duration of observation was 2.6 h (SD: 1.8 h) and 120.0 h (SD: 0.2 h), respectively. During that time, a total of 395,213 data points were included. After excluding the capnometry values < 2 mmHg, 392,808 data points remained. The mean number of data points per operating room patient was 3087.1 (SD: 3117.0); in the ICU, it was 4324.4 (SD: 2880.6). Of these, the four human experts annotated a total of 11,699 data points as artifacts, including the data points annotated by two experts evaluating each patient. In 2,891 data points, consensus was met by the first step, leaving 5,917 data points (50.6%) for the third expert's decision. In 2,276 of those cases (38.5%), he decided that the data point was an artifact, resulting in 5,167 annotated artifacts (1.3% of the data points). Splitting the data into a training set (80%) and a test set (20%) resulted in 310,085 instances without artifacts and 4,162 instances (1.33%) with artifacts in the training set. In the test set, there were 77,557 instances without artifacts and 1,005 instances with artifacts (1.28%).

The application of the artifact detection algorithms resulted in a large variation in annotated artifacts. For example, the interquartile range method resulted in 6,196 artifacts annotated, while the local outlier factor annotated only 1,189 data points as artifacts. Details can be found in Table 2. Data describing the long short-term memory neural network are missing in Table 2 due to the dataset split.

The hypothesis that the neural network showed significantly higher sensitivities than the classic methods was found to be true for the following vital signs: heart rate (ICU: 33.6%, 95% CI: 33.1–44.6), systolic invasive blood pressure (both in the operating room (62.2%,

95% CI: 57.5–71.9) and ICU (60.7%, 95% CI: 57.3–71.8), and temperature in the operating room (76.1%, 95% CI: 63.6–89.7). Specificity was very similar in all methods. As expected, the interquartile range and z-value performed very similar, and the data were equally distributed. The best-performing methods are summarized in Tables 3 and 4.

The specificity was above 90% for all detection methods and all vital signs. However, the sensitivity was low for cutoff, z-value, interquartile range, and local outlier factor, with only the z-value for saturation in the operating room reaching 88.9%. All other sensitivity values were less than 80%, with the local outlier factor not exceeding 10% of the sensitivity. An example of the neural network's performance can be found in Figure 1.

A comparison of the performance across methods revealed significant differences between vital signs, methods, and clinical locations. For example, for heart rate in the ICU, the long short-term memory neural network showed a significantly higher sensitivity of 33.6% and specificity of 99.2%, whereas the interquartile range showed 19.5%/99.4%, the z-value performed similarly (25.3%/99.6%), and the cutoff showed a sensitivity of only 3.8%, with a specificity of 100%. By contrast, the cutoff showed better results for invasive mean arterial pressure (MAP) in the operating room (sensitivity: 74.9%, specificity: 100%) but not in the ICU (9.3%/100%). Details, including 95% confidence intervals, can be found in Table 3, showing performance in data from the operating room, and in Table 4, showing performance in data from the ICU.

To show the performance of different thresholds when using the interquartile range and z-value, a sensitivity analysis was conducted: all threshold values from 0.5 to 5 were tested, as well as different cutoff values. The results showed that using values other than those previously described resulted in better sensitivity, while specificity stayed at an acceptable level. For example, the second threshold level (values worrying the treating healthcare professionals) resulted in 75% sensitivity for invasive MAP. However, specificity decreased to 92%, while those originally used showed 39.8%/100%. This trend was seen in all used

thresholds: increased sensitivity led to rapidly decreasing specificity. All calculations can be found in Table 5 and in the supplement (https://links.lww.com/ALN/D489).

All ROC area under the curve (AUC) values were above 0.61, with most exceeding 0.85. For example, applying the z-value to MAP resulted in an ROC-AUC of 0.88, while applying the interquartile range to CO_2 resulted in an AUC of 0.96. The resulting ROC curves, including all AUC values, can be found in the supplementary materials (https://links.lww.com/ALN/D489).

- Discussion

In the present study, we found that artifact filtering methods performed differently both in terms of specific vital signs and clinical context of intraoperative versus intensive care unit. No one method was found to be consistently superior across different vital signs and clinical contexts. Compared to human experts annotating artifacts retrospectively,[8,9,36] the methods of interquartile range, z-value, and cutoff filters showed high specificity but only intermediate sensitivity; the local outlier factor had a sensitivity below 10%. By contrast, a specially trained, long short-term memory neural network showed higher sensitivity values, while specificity remained as high as the other methods. Narrowing the thresholds of the cutoff filter in a sensitivity analysis also increased sensitivity; however, specificity decreased rapidly. The thoughtful selection of artifact detection methods for each clinical parameter is important. For specific clinical parameters, the use of neural networks demonstrated higher artifact filtering performance.

Artifact filtering is of the utmost importance, as it has the potential to alter scientific results.[8,12] In the vast majority of anesthesiologic and intensive care publications to date, only basic methods of artifact detection in recordings of continuous vital parameters have been reported.[1–3,7] Some studies have not described any detail of artifact detection at all. However, a broad variety of artifact detection algorithms and highly specialized neural networks have

been published for artifact detection in retrospective data.[37–39] In the present study, it was shown that different artifact detection methods perform differently on each vital sign.

Blood pressure is the focus of many perioperative and critical care research efforts, with both MAP and systolic pressure being reported.[1,3,5,7] No single method performed best for invasive MAP: In the operating room, the cutoff method performed best, while in the ICU, the neural network performed best. The sensitivity analysis showed that narrowing the limits rapidly increased sensitivity, with a drop in specificity. This further emphasizes the relevance of choosing the right algorithm with the right threshold. This is especially relevant when looking at pathological states, such as intraoperative hypotension or hypothermia: erroneously flagging vital signals as artifacts would lead to the exclusion of relevant information.

The same importance of choosing the right method for the right parameter was seen for heart rate, for which all algorithms showed a sensitivity of less than 40% with large differences between the ICU and the operating room. Although the sensitivity analysis showed a tendency toward increased sensitivity, specificity dropped rapidly. The most probable cause for the low sensitivities is rapid changes in heart rate: electrocautery in the operating room and movement or arrhythmia in the ICU. These artifacts can easily be detected when using the pulse rate from the oxygen saturation—a signal not available to the artifact detection methods.

Data on artifact filtering for vital signs other than heart rate and blood pressure are sparse. The current data demonstrate that each method performed differently for temperature, pulse oximetry, and capnometry, with regard to sensitivity and specificity. In contrast to heart rate and blood pressure, z-value and interquartile range performed well, while the neural network showed mixed results.

As with blood pressure and heart rate, choosing the right threshold—especially for cut-off filters—is an important topic. Although sensitivity was increased by changing the thresholds, specificity dropped for temperature and capnometry. Only for pulse oximetry did the

specificity remain stable, which can most probably be attributed to the special distribution of values.

Interestingly, all algorithms performed worse in the ICU than in the operating room. There are multiple potential explanations for this difference, with the decreased resolution of data in the ICU being the most important. For example, although an artifact due to blood sampling can easily be detected in the arterial blood pressure signal with a 15-second resolution, it will diminish at a 15-minute resolution.

Due to the similar statistical approaches used for the interquartile range and z-value, similar results were expected. Nevertheless, for datasets with a "low resolution" vital sign, such as the noninvasive blood pressure in this study, the z-value often performed better compared to the interquartile range. Generally, the local outlier factor performed poorly throughout all vital signs, with sensitivity never exceeding 15%. One conclusion may be that the use of the local outlier factor in anesthetic data is questionable, contrary to other datasets.[40] The topic of data granularity itself could not be studied in this project, as the data resolution was evenly distributed. Further research is necessary, especially on the effects of data granularity on artifact filtering algorithms and the characterizations of different vital signs, which are most often subsequently used for statistical analysis. As the collection of high-resolution vital signs, including waveforms, becomes increasingly popular, the problem of low granular data will improve —at least in perioperative data.

Limitations

One significant limitation of this study is the use of retrospective EHR data as the foundation for creating the human reference standard rather than real-time point of care annotation. This possibly explains why the rate of artifacts marked by the reviewers (1.3%) was lower than in some previous studies reporting up to 14%.[8,9,39] Using retrospective data can lead to missing artifacts. For example, the wrong elevation of a blood pressure transducer cannot be identified in retrospective data, patient movement cannot be seen, and the dislocation of a saturation

sensor can only be assumed. However, as the main objective was a comparison of artifact filtering methods for use in retrospective data, we decided to choose a retrospective approach for human experts as well.

The neural network used has some limitations. Of importance is the fact that the tested network was trained with the human reference standard and therefore could at best predict the reference standard. External validation to prove generalizability is needed before using the model in other centers. In this case, approximately 450 person-hours were needed to annotate all 106 patients, including the majority vote, for artifacts where two reviewers disagreed. Another limitation is that humans, classic artifact detection algorithms, and neural networks may not be comparable. Although the neural network is trained with actual reference standard data, such as human expert reviewer annotations, the classic algorithms do not have access to individually annotated data. Classic algorithms are based on thresholds, population distributions, or mathematical transformations. Humans, by contrast, can utilize much more information to filter artifacts than is available for long short-term memory neural networks. The use of electrocautery or the movement of cables is obvious to the human eye but difficult to learn for long short-term memory neural networks. Neural network training has the potential to overfit, which was decreased using a randomly split internal test dataset. As shown in the supplementary materials (https://links.lww.com/ALN/D489), the neural network's calibration plots had poor calibration for certain variables, such as pulse oximetry values and temperature, both in the ICU and the operating room, whereas they showed good performance for others, such as systolic blood pressure of heart rate, in both clinical settings. This indicates that there may be a relevant risk of decision errors in particular risk ranges. The use of human experts is a significant limitation of this study. Although all experts had experience in clinical anesthesia and intensive care, incorrect filtering was always possible. To address this concern, every data point was independently evaluated by two experts. The distribution of patient data to two of the four experts was performed randomly. For cases in

which these two experts diverged, a third expert decided with a majority vote. The number of data points in which this majority vote was necessary was high (50.6% of annotated data points), leading to the concern that the analysis relied on an imperfect reference standard. The use of three independent experts is the maximum previously described in the literature, leading to the assumption that the reference standard used is the best possible in that setting. [8,39] Following Walsh, we conducted a naïve analysis accepting that results may be underestimated or overestimated, while other methods seemed impractical in this case. [41]

A further limitation is the single-center data source, which has limited generalizability. The highly standardized way in which monitoring devices are produced and used is encouraging, but we cannot exclude systematic errors, such as those described previously. [42,43] In addition, many commercially available EHRs record only intraoperative physiologic data every 60 seconds, calling into question the validity of our findings in these datasets. Validation of the neural network used on external, completely new data is essential to establish the potential value of the network on broad patient populations.

- Conclusion

The use of simple, universal physiologic artifact detection methods seems inferior to vital sign-specific artifact detection algorithms. Commonly used artifact detection methods performed very differently when tested for different vital signs and for different settings (ICU vs. operating room). Using a pretrained neural network for artifact filtering in retrospective data may be a possible additional valid option as an artifact detection method, although performance may be worse in different datasets and potential overfitting is a an important limitation.

Supplemental Digital Content

Supplement 1: ROC curves, calibration plots, STARD checklist, https://links.lww.com/ALN/D489

Supplement 2: Jupyter notebook of Python code, https://links.lww.com/ALN/D490

Supplement 3: Jupyter notebook with a simplified LSTM's code for learning, https://links.lww.com/ALN/D491

Supplement 4: Supplemental Table 1: Artifact detection algorithms' performance in combined data; Supplemental Table 2: Sensitivity analysis of cut-off method, https://links.lww.com/ALN/D492

Supplement 5: Description of the human reviewers' training, https://links.lww.com/ALN/D493

- References

1. Walsh M, Devereaux PJ, Garg AX, Kurz A, Turan A, Rodseth RN, Cywinski J, Thabane L, Sessler DI: Relationship between intraoperative mean arterial pressure and clinical outcomes after noncardiac surgery: toward an empirical definition of hypotension. Anesthesiology 2013; 119:507–15
2. Sun LY, Wijeysundera DN, Tait GA, Beattie WS: Association of intraoperative hypotension with acute kidney injury after elective noncardiac surgery. Anesthesiology 2015; 123:515–23
3. Bijker JB, Persoon S, Peelen LM, Moons KGM, Kalkman CJ, Kappelle LJ, Klei WA van: Intraoperative hypotension and perioperative ischemic stroke after general surgery: a nested case-control study. Anesthesiology 2012; 116:658–64
4. Gregory A, Stapelfeldt WH, Khanna AK, Smischney NJ, Boero IJ, Chen Q, Stevens M, Shaw AD: Intraoperative Hypotension Is Associated With Adverse Clinical Outcomes After Noncardiac Surgery. Anesth Analg 2021; 132:1654–65
5. Wesselink EM, Kappen TH, Torn HM, Slooter AJC, Klei WA van: Intraoperative hypotension and the risk of postoperative adverse outcomes: a systematic review. British Journal of Anaesthesia 2018; 121:706–21
6. Johnson AEW, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, Moody B, Szolovits P, Anthony Celi L, Mark RG: MIMIC-III, a freely accessible critical care database. Sci Data 2016; 3:160035
7. Salmasi V, Maheshwari K, Yang D, Mascha EJ, Singh A, Sessler DI, Kurz A: Relationship between Intraoperative Hypotension, Defined by Either Reduction from Baseline or Absolute Thresholds, and Acute Kidney and Myocardial Injury after Noncardiac Surgery: A Retrospective Cohort Analysis. Anesthesiology 2017; 126:47–65
8. Hoorweg A, Pasma W, Wolfswinkel L van, Graaff JD de: Incidence of Artifacts and Deviating Values in Research Data Obtained from an Anesthesia Information

Management System in Children. Anesthesiology 2018 doi:10.1097/ALN.0000000000001895

9. Kool NP, Waes JAR van, Bijker JB, Peelen LM, Wolfswinkel L van, Graaff JC de, Klei WA van: Artifacts in research data obtained from an anesthesia information and management system. Can J Anaesth 2012; 59:833–41
10. Takla G, Petre JH, Doyle DJ, Horibe M, Gopakumaran B: The problem of artifacts in patient monitor data during surgery: a clinical and methodological review. Anesth Analg 2006; 103:1196–204
11. Hoare SW, Beatty PC: Automatic artifact identification in anaesthesia patient record keeping: a comparison of techniques. Med Eng Phys 2000; 22:547–53
12. Pasma W, Peelen LM, Buuren S van, Klei WA van, Graaff JC de: Artifact Processing Methods Influence on Intraoperative Hypotension Quantification and Outcome Effect Estimates. Anesthesiology 2020; 132:723–37
13. Chen L, Dubrawski A, Wang D, Fiterau M, Guillame-Bert M, Bose E, Kaynar AM, Wallace DJ, Guttendorf J, Clermont G, Pinsky MR, Hravnak M: Using Supervised Machine Learning to Classify Real Alerts and Artifact in Online Multisignal Vital Sign Monitoring Data. Crit Care Med 2016; 44:e456-463
14. Du CH, Glick D, Tung A: Error-checking intraoperative arterial line blood pressures. J Clin Monit Comput 2019; 33:407–12
15. Hravnak M, Chen L, Dubrawski A, Bose E, Clermont G, Pinsky MR: Real alerts and artifact classification in archived multi-signal vital sign monitoring data: implications for mining big data. J Clin Monit Comput 2016; 30:875–88
16. Pao Y: Adaptive pattern recognition and neural networks. United States, Reading, MA (US); Addison-Wesley Publishing Co., Inc., 1989 at <https://www.osti.gov/biblio/5238955>

17. Ripley BD: Pattern Recognition and Neural Networks. Cambridge, Cambridge University Press, 1996 doi:10.1017/CBO9780511812651
18. Van Rossum G, Drake Jr FL: Python tutorial. Centrum voor Wiskunde en Informatica Amsterdam, The Netherlands, 1995
19. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, Lijmer JG, Moher D, Rennie D, Vet HCW de, Kressel HY, Rifai N, Golub RM, Altman DG, Hooft L, Korevaar DA, Cohen JF, STARD Group: STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. BMJ 2015; 351:h5527
20. Breunig MM, Kriegel H-P, Ng RT, Sander J: LOF: identifying density-based local outliers. SIGMOD Rec 2000; 29:93–104
21. Kingma DP, Ba J: Adam: A Method for Stochastic Optimization 2017 doi:10.48550/arXiv.1412.6980
22. McKinney W: Data Structures for Statistical Computing in Python, Proceedings of the 9th Python in Science Conference. Edited by Walt S van der, Millman J. 2010, pp 56–61 doi:10.25080/Majora-92bf1922-00a
23. Harris CR, Millman KJ, Walt SJ van der, Gommers R, Virtanen P, Cournapeau D, Wieser E, Taylor J, Berg S, Smith NJ, Kern R, Picus M, Hoyer S, Kerkwijk MH van, Brett M, Haldane A, Río JF del, Wiebe M, Peterson P, Gérard-Marchant P, Sheppard K, Reddy T, Weckesser W, Abbasi H, Gohlke C, Oliphant TE: Array programming with NumPy. Nature 2020; 585:357–62
24. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay É: Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 2011; 12:2825–30
25. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, Walt SJ van der, Brett M, Wilson J, Millman KJ,

Mayorov N, Nelson ARJ, Jones E, Kern R, Larson E, Carey CJ, Polat İ, Feng Y, Moore EW, VanderPlas J, Laxalde D, Perktold J, Cimrman R, Henriksen I, Quintero EA, Harris CR, et al.: SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods 2020; 17:261–72

26. Wilson EB: Probable Inference, the Law of Succession, and Statistical Inference. Journal of the American Statistical Association 1927; 22:209–12
27. Altman D: Statistics with Confidence: Confidence Intervals and Statistical Guidelines, 2nd edition. Edited by Machin D, Bryant T, Gardner M. BMJ Books, 2013
28. Cumming G, Finch S: Inference by Eye: Confidence Intervals and How to Read Pictures of Data. American Psychologist 2005; 60:170–80
29. Liu W, Bretz F, Cortina-Borja M: Reference range: Which statistical intervals to use? Stat Methods Med Res 2021; 30:523–34
30. Menacer S, Claessens Y-E, Meune C, Elfassi Y, Wakim C, Gauthier L, Fortun M, Goudot F-X, Dehoux M, Lefèvre G, Chenevier-Gobeaux C: Reference range values of troponin measured by sensitive assays in elderly patients without any cardiac signs/symptoms. Clin Chim Acta 2013; 417:45–7
31. Roshan D, Ferguson J, Pedlar CR, Simpkin A, Wyns W, Sullivan F, Newell J: A comparison of methods to generate adaptive reference ranges in longitudinal monitoring. PLoS One 2021; 16:e0247338
32. Brouwers S, Sudano I, Kokubo Y, Sulaica EM: Arterial hypertension. Lancet 2021; 398:249–61
33. Bos LD, Martin-Loeches I, Schultz MJ: ARDS: challenges in patient care and frontiers in research. Eur Respir Rev 2018; 27:170107
34. ARDS Definition Task Force, Ranieri VM, Rubenfeld GD, Thompson BT, Ferguson ND, Caldwell E, Fan E, Camporota L, Slutsky AS: Acute respiratory distress syndrome: the Berlin Definition. JAMA 2012; 307:2526–33

35. Ibanez B, James S, Agewall S, Antunes MJ, Bucciarelli-Ducci C, Bueno H, Caforio ALP, Crea F, Goudevenos JA, Halvorsen S, Hindricks G, Kastrati A, Lenzen MJ, Prescott E, Roffi M, Valgimigli M, Varenhorst C, Vranckx P, Widimský P, ESC Scientific Document Group: 2017 ESC Guidelines for the management of acute myocardial infarction in patients presenting with ST-segment elevation: The Task Force for the management of acute myocardial infarction in patients presenting with ST-segment elevation of the European Society of Cardiology (ESC). European Heart Journal 2018; 39:119–77
36. Hravnak M, Chen L, Bose E, Fiterau M, Guillame-Bert M, Dubrawski A, Clermont G, Pinsky M: Artifact Patterns in Continuous Noninvasive Monitoring of Patients. Intensive Care Med 2013; 39:S405
37. Simpao AF, Nelson O, Ahumada LM: Automated anesthesia artifact analysis: can machines be trained to take out the garbage? J Clin Monit Comput 2021; 35:225–7
38. Hashimoto DA, Witkowski E, Gao L, Meireles O, Rosman G: Artificial Intelligence in Anesthesiology: Current Techniques, Clinical Applications, and Limitations. Anesthesiology 2020; 132:379–94
39. Pasma W, Wesselink EM, Buuren S van, Graaff JC de, Klei WA van: Artifacts annotations in anesthesia blood pressure data by man and machine. J Clin Monit Comput 2021; 35:259–67
40. He T, Zhou Q, Zou Y: Automatic Detection of Age-Related Macular Degeneration Based on Deep Learning and Local Outlier Factor Algorithm. Diagnostics (Basel) 2022; 12:532
41. Walsh T: Fuzzy gold standards: Approaches to handling an imperfect reference standard. Journal of Dentistry 2018; 74:S47–9
42. Fawzy A, Wu TD, Wang K, Robinson ML, Farha J, Bradke A, Golden SH, Xu Y, Garibaldi BT: Racial and Ethnic Discrepancy in Pulse Oximetry and Delayed Identification of Treatment Eligibility Among Patients With COVID-19. JAMA Internal Medicine 2022; 182:730–8

43. Bothe TL, Bilo G, Parati G, Haberl R, Pilz N, Patzak A: Impact of oscillometric measurement artefacts in ambulatory blood pressure monitoring on estimates of average blood pressure and of its variability: a pilot study. J Hypertens 2023; 41:140–9

- Figures Legends

Figure 1: Example of the neural network performance compared to the human experts in a sample of heart rates from one specific patient (example signal: normalized heart rate). By converting the minimal heart rate to -1 and the maximal heart rate to 1, the performance of the neural network is facilitated, while intervals between values of heart rate remain the same.

- Table 1: Demographic Details

	OR (n = 53)	ICU (n = 53)
Age (Quartile)	59.0 (42.0, 68.0)	60.0 (49.0, 67.0)
BMI (Quartile)	1.9 (1.7, 2.0)	1.9 (1.8, 2.1)
Female gender (%)	20 (37.7%)	21 (39.6%)
Length of stay [hours] (Quartile)	2.5 (1.9, 3.6)	335.9 (237.2, 551.8)
ASA classification n (%)		
ASA 1	8 (15.1%)	nA
ASA 2	24 (45.3%)	nA
ASA 3	20 (37.7%)	nA
ASA 4	1 (1.9%)	nA
Surgical Specialty n (%)		
General surgery, Gynecology, Urology	24 (45.3%)	nA
Cardiothoracic, Vascular	6 (11.3%)	nA
ENT, Maxillofacial, Dermatology	6 (11.3%)	nA
Ortho, Trauma, Ophthalmology	11 (20.8%)	nA
Neurosurgery	5 (9.4%)	nA
Robotic surgery	1 (1.9%)	nA
Type of ICU admission n (%)		
Planned	nA	23 (43.4%)
Unplanned	nA	30 (56.6%)
Surgical status n (%)		
Planned surgery	nA	11 (20.8%)
Urgent surgery	nA	29 (54.7%)
No surgery	nA	13 (24.5%)

Table 1 shows basic demographic details of the included patients. OR: operating room, n = number, ASA: ASA physical status, Ortho: orthopedics, CPR: cardiopulmonary resuscitation

- Table 2: Number of annotated artifacts

Vital sign	Data points			Human			IQR			z-value			LOF			Cutoff		
	ICU	OR	Total	ICU	OR	Total	ICU	OR	Total	ICU	OR	Total	ICU	OR	Total	ICU	OR	Total
Heart rate	38,307	34,998	73,305	655	990	1,645	344	597	941	329	470	799	115	50	165	25	22	47
SpO2	38,247	35,066	73,313	88	36	124	1,161	951	2,112	455	416	871	82	24	106	0	0	0
Blood pressure																		
Systolic, invasive	36,927	15,451	52,378	487	423	910	157	374	531	257	240	497	106	59	165	57	252	309
MAP, invasive	36,901	15,414	52,315	440	382	822	182	348	530	261	264	525	107	56	163	41	287	328
Diastolic, invasive	36,887	15,411	52,298	464	375	839	216	398	614	310	256	566	97	49	146	42	90	132
Systolic, noninvasive	611	3,017	3,628	0	7	7	0	18	18	5	24	29	24	3	27	0	0	0
MAP, noninvasive	612	3,017	3,629	2	7	9	0	32	32	7	28	35	24	3	27	0	0	0
Diastolic, noninvasive	611	3,017	3,628	0	11	11	0	27	27	7	32	39	24	3	27	0	0	0
Capnometry	18,487	34,903	53,390	29	523	552	235	1,045	1,142	148	644	792	207	37	244	1,954	1,800	3,754
Temperature	21,604	3,320	24,924	210	38	248	227	22.0	249	200	18	218	119	0	119	94	10	104
Total	229,194	163,614	392,808	2,375	2,325	5,167	2,866	4,387	6,196	2,308	2,862	4,371	1,020	334	1,189	2,238	701	4,674

Table 2 shows details of the included data points as well as the number of annotated artifacts by detection method. OR: operating room, IQR: interquartile range, LOF: local outlier factor, SpO2: peripheral oxygen saturation, MAP: mean arterial pressure

Table 3: ***Performance of the Algorithms in the Operating Room***

Vital sign	Method	Sensitivity (CI95)	Specificity (CI95)	PPV (CI95)	NPV (CI95)
CO2	**IQR ***	**72.5 (68.5, 76.1)**	**98.1 (97.9, 98.2)**	36.3 (33.4, 39.2)	99.6 (99.5, 99.6)
	Z-value	**70.6 (66.5, 74.3)**	**99.2 (99.1, 99.3)**	57.3 (53.4, 61.1)	99.6 (99.5, 99.6)
	Local outlier factor	4.2 (2.8, 6.3)	100 (99.9, 100.0)	59.5 (43.5, 73.7)	98.6 (98.4, 98.7)
	Cutoff	1.5 (0.8, 3.0)	94.8 (94.5, 95.0)	0.4 (0.2, 0.9)	98.4 (98.3, 98.6)
	Neural Network	**71.1 (61.5, 79.3)**	**99.6 (99.6, 99.8)**	64.6 (63.4, 81.6)	99.7 (99.5, 99.8)
Heart rate	IQR	34.2 (31.4, 37.3)	99.2 (99.1, 99.3)	56.8 (52.8, 60.7)	98.1 (98.0, 98.2)
	Z-value	29.7 (26.9, 32.6)	99.5 (99.4, 99.6)	62.6 (58.1, 66.8)	98 (97.8, 98.1)
	Local outlier factor	4 (3.0, 5.5)	100 (99.9, 100.0)	80 (67.0, 88.8)	97.3 (97.1, 97.4)
	Cutoff	2.2 (1.5, 3.3)	100 (100.0, 100.0)	100 (85.1, 100.0)	97.2 (97.1, 97.4)
	Neural Network *	39.5 (33.7, 44.8)	98.9 (98.9, 99.2)	42.9 (38.5, 50.6)	98.7 (98.6, 99.0)
SpO2	**IQR**	**72.2 (56.0, 84.2)**	**97.4 (97.2, 97.5)**	2.7 (1.9, 4.0)	100 (99.9, 100.0)
	Z-value *	**88.9 (74.7, 95.6)**	**98.9 (98.8, 99.0)**	7.7 (5.5, 10.7)	100 (100.0, 100.0)

	Local outlier factor	8.3 (2.9, 21.8)	99.9 (99.9, 100.0)	12.5 (4.3, 31.0)	99.9 (99.9, 99.9)
	Cutoff	0 (0.0, 9.6)	100 (100.0, 100.0)	(nan, nan)	99.9 (99.9, 99.9)
	Neural Network	13.6 (8.3, 45.8)	100 (100.0, 100.0)	100 (100.0, 100.0)	99.9 (99.9, 100.0)
Temperature	IQR	23.7 (13.0, 39.2)	99.6 (99.3, 99.8)	40.9 (23.3, 61.3)	99.1 (98.7, 99.4)
	Z-value	31.6 (19.1, 47.5)	99.8 (99.6, 99.9)	66.7 (43.7, 83.7)	99.2 (98.8, 99.5)
	Local outlier factor	0 (0.0, 9.2)	100 (99.9, 100.0)	(nan, nan)	98.9 (98.4, 99.2)
	Cutoff	26.3 (15.0, 42.0)	100 (99.9, 100.0)	100 (72.2, 100.0)	99.2 (98.8, 99.4)
	Neural Network *	**76.1 (63.6, 89.7)**	**99.9 (99.8, 100.0)**	74.5 (66.7, 90.9)	99.9 (99.8, 100.0)
Systolic, IBP	IQR	49.4 (44.7, 54.2)	98.9 (98.7, 99.1)	55.9 (50.8, 60.8)	98.6 (98.4, 98.8)
	Z-value	36.9 (32.4, 41.6)	99.4 (99.3, 99.5)	65 (58.8, 70.8)	98.2 (98.0, 98.4)
	Local outlier factor	13.5 (10.5, 17.1)	100 (100.0, 100.0)	96.6 (88.5, 99.1)	97.6 (97.4, 97.9)
	Cutoff	59.6 (54.8, 64.1)	100 (100.0, 100.0)	100 (98.5, 100.0)	98.9 (98.7, 99.0)
	Neural Network *	62.2 (57.5, 71.9)	99.9 (99.8, 100.0)	88.7 (84.4, 95.0)	99.5 (99.4, 99.6)

Diastolic, IBP	IQR *	49.9 (44.8, 54.9)	98.6 (98.4, 98.8)	47 (42.1, 51.9)	98.7 (98.6, 98.9)
	Z-value	46.9 (41.9, 52.0)	99.5 (99.3, 99.6)	68.8 (62.8, 74.1)	98.7 (98.5, 98.9)
	Local outlier factor	12.5 (9.6, 16.3)	100 (100.0, 100.0)	95.9 (86.3, 98.9)	97.9 (97.6, 98.1)
	Cutoff	24 (20.0, 28.6)	100 (100.0, 100.0)	100 (95.9, 100.0)	98.1 (97.9, 98.3)
	Neural Network	33.3 (21.5, 41.3)	99.9 (99.8, 100.0)	75.8 (57.9, 87.5)	99.4 (99.1, 99.5)
MAP, IBP	IQR	47.1 (42.2, 52.1)	98.9 (98.7, 99.0)	51.7 (46.5, 56.9)	98.7 (98.5, 98.8)
	Z-value	46.1 (41.1, 51.1)	99.4 (99.3, 99.5)	66.7 (60.8, 72.1)	98.6 (98.4, 98.8)
	Local outlier factor	14.1 (11.0, 18.0)	100 (100.0, 100.0)	96.4 (87.9, 99.0)	97.9 (97.6, 98.1)
	Cutoff *	**74.9 (70.3, 79.0)**	**100 (100.0, 100.0)**	99.7 (98.1, 99.9)	99.4 (99.2, 99.5)
	Neural Network	54.4 (38.4, 56.7)	99.8 (99.7, 99.9)	75.6 (69.8, 88.7)	99.4 (99.1, 99.4)
Systolic, NIBP	IQR	0 (0.0, 35.4)	99.4 (99.1, 99.6)	0 (0.0, 17.6)	99.8 (99.5, 99.9)
	Z-value*	28.6 (8.2, 64.1)	99.3 (98.9, 99.5)	8.3 (2.3, 25.8)	99.8 (99.6, 99.9)
	Local outlier factor	0 (0.0, 35.4)	99.9 (99.7, 100.0)	0 (0.0, 56.1)	99.8 (99.5, 99.9)
	Cutoff	0 (0.0, 35.4)	100 (99.9, 100.0)	(nan, nan)	99.8 (99.5, 99.9)
	Neural Network				

Diastolic, NIBP	IQR	0 (0.0, 25.9)	99.1 (98.7, 99.4)	0 (0.0, 12.5)	99.6 (99.3, 99.8)
	Z-value*	36.4 (15.2, 64.6)	99.1 (98.7, 99.4)	12.5 (5.0, 28.1)	99.8 (99.5, 99.9)
	Local outlier factor	0 (0.0, 25.9)	99.9 (99.7, 100.0)	0 (0.0, 56.1)	99.6 (99.3, 99.8)
	Cutoff	0 (0.0, 25.9)	100 (99.9, 100.0)	(nan, nan)	99.6 (99.3, 99.8)
	Neural Network				
MAP, NIBP	IQR *	42.9 (15.8, 75.0)	99 (98.6, 99.3)	9.4 (3.2, 24.2)	99.9 (99.7, 99.9)
	Z-value *	42.9 (15.8, 75.0)	99.2 (98.8, 99.4)	10.7 (3.7, 27.2)	99.9 (99.7, 99.9)
	Local outlier factor	0 (0.0, 35.4)	99.9 (99.7, 100.0)	0 (0.0, 56.1)	99.8 (99.5, 99.9)
	Cutoff	0 (0.0, 35.4)	100 (99.9, 100.0)	(nan, nan)	99.8 (99.5, 99.9)
	Neural Network				

Table 3 shows the sensitivity, specificity, positive predictive value, and negative predictive value of all artifact detection algorithms separated in the operating room. CI95: 95% confidence interval, OR: operating room, PPV: positive predictive value, NPV: negative predictive value, MAP: mean arterial pressure, Systolic: systolic blood pressure, Diastolic: diastolic blood pressure, IBP: invasive blood pressure, NIBP: non-invasive blood pressure, IQR: interquartile range, Neural Network: long-short term memory (machine learning algorithm). Note that too little information was available to train the neural net for non-invasive blood pressure. All cells with a sensitivity above 70% and a specificity above 95% are marked in bold. Asterisks mark methods with the highest sensitivity, specificity was >97% in all marked methods.

- Table 4: Performance of the algorithms in Intensive Care Unit

Vital sign	Method	Sensitivity (CI95)	Specificity (CI95)	PPV (CI95)	NPV (CI95)
CO2	IQR	51.7 (34.4, 68.6)	98.8 (98.6, 99.0)	6.4 (3.9, 10.3)	99.9 (99.9, 100.0)
	Z-value	62.1 (44.0, 77.3)	99.3 (99.2, 99.4)	12.2 (7.8, 18.4)	99.9 (99.9, 100.0)
	Local outlier factor	10.3 (3.6, 26.4)	98.9 (98.7, 99.0)	1.4 (0.5, 4.2)	99.9 (99.8, 99.9)
	Cutoff	6.9 (1.9, 22.0)	89.4 (89.0, 89.9)	0.1 (0.0, 0.4)	99.8 (99.8, 99.9)
	Neural Network *	**72.6 (61.4, 79.2)**	**99.7 (99.6, 99.8)**	74.5 (63.4, 81.8)	99.6 (99.5, 99.8)
Heart rate	IQR	19.5 (16.7, 22.8)	99.4 (99.3, 99.5)	37.2 (32.3, 42.4)	98.6 (98.5, 98.7)
	Z-value	25.3 (22.2, 28.8)	99.6 (99.5, 99.6)	50.5 (45.1, 55.8)	98.7 (98.6, 98.8)
	Local outlier factor	6.6 (4.9, 8.7)	99.8 (99.8, 99.8)	37.4 (29.1, 46.5)	98.4 (98.3, 98.5)
	Cutoff	3.8 (2.6, 5.6)	100 (100.0, 100.0)	100 (86.7, 100.0)	98.4 (98.2, 98.5)
	Neural Network *	33.6 (33.2, 44.6)	99.2 (98.9, 99.2)	47.6 (38.6, 50.8)	98.6 (98.6, 99.0)
SpO2	IQR	64.8 (54.4, 73.9)	97.1 (96.9, 97.3)	4.9 (3.8, 6.3)	99.9 (99.9, 99.9)
	Z-value *	**73.9 (63.8, 81.9)**	**99 (98.9, 99.1)**	14.3 (11.4, 17.8)	99.9 (99.9, 100.0)
	Local outlier factor	0 (0.0, 4.2)	99.8 (99.7, 99.8)	0 (0.0, 4.5)	99.8 (99.7, 99.8)
	Cutoff	0 (0.0, 4.2)	100 (100.0, 100.0)	(nan, nan)	99.8 (99.7, 99.8)

	Neural Network	18.2 (9.1-45.0)	100 (100.0, 100.0)	100 (100.0, 100.0)	100 (99.9, 100.0)
Temperature	**IQR ***	**71.9 (65.5, 77.5)**	**99.6 (99.6, 99.7)**	66.5 (60.2, 72.3)	99.7 (99.6, 99.8)
	Z-value	68.1 (61.5, 74.0)	99.7 (99.7, 99.8)	71.5 (64.9, 77.3)	99.7 (99.6, 99.8)
	Local outlier factor	1.4 (0.5, 4.1)	99.5 (99.4, 99.5)	2.5 (0.9, 7.2)	99 (98.9, 99.2)
	Cutoff	42.4 (35.9, 49.1)	100 (99.9, 100.0)	94.7 (88.1, 97.7)	99.4 (99.3, 99.5)
	Neural Network	66.7 (37.5, 83.3)	100 (100.0, 100.0)	100 (100.0, 100.0)	99.7 (99.4, 99.9)
Systolic, IBP	IQR	29.8 (25.9, 34.0)	100 (99.9, 100.0)	92.4 (87.1, 95.6)	99.1 (99.0, 99.2)
	Z-value	37 (32.8, 41.3)	99.8 (99.7, 99.8)	70 (64.2, 75.3)	99.2 (99.1, 99.3)
	Local outlier factor	7.6 (5.6, 10.3)	99.8 (99.8, 99.9)	34.9 (26.5, 44.4)	98.8 (98.7, 98.9)
	Cutoff	11.7 (9.1, 14.9)	100 (100.0, 100.0)	100 (93.7, 100.0)	98.8 (98.7, 98.9)
	Neural Network *	60.7 (57.3, 71.8)	99.9 (99.8, 100.0)	88.3 (84.1, 95.0)	99.5 (99.4, 99.6)
Diastolic, IBP	IQR	38.1 (33.8, 42.6)	99.9 (99.9, 99.9)	81.9 (76.3, 86.5)	99.2 (99.1, 99.3)
	Z-value *	41.6 (37.2, 46.1)	99.7 (99.6, 99.7)	62.3 (56.7, 67.5)	99.3 (99.2, 99.3)
	Local outlier factor	6.2 (4.4, 8.8)	99.8 (99.8, 99.9)	29.9 (21.7, 39.6)	98.8 (98.7, 98.9)
	Cutoff	9.1 (6.8, 12.0)	100 (100.0, 100.0)	100 (91.6, 100.0)	98.9 (98.7, 99.0)
	Neural Network	28.9 (21.4, 42.1)	99.9 (99.8-100.0)	74.3 (58.1, 87.5)	99.2 (99.1, 99.5)

MAP, IBP	IQR	34.8 (30.5, 39.3)	99.9 (99.9, 99.9)	84.1 (78.1, 88.7)	99.2 (99.1, 99.3)
	Z-value	40 (35.5, 44.6)	99.8 (99.7, 99.8)	67.4 (61.5, 72.8)	99.3 (99.2, 99.4)
	Local outlier factor	9.1 (6.7, 12.1)	99.8 (99.8, 99.9)	37.4 (28.8, 46.8)	98.9 (98.8, 99.0)
	Cutoff	9.3 (6.9, 12.4)	100 (100.0, 100.0)	100 (91.4, 100.0)	98.9 (98.8, 99.0)
	Neural Network *	51.5 (38.9, 57.7)	99.8 (99.7, 99.9)	80.7 (69.7, 88.7)	99.2 (99.1, 99.4)
Systolic, NIBP	IQR	(nan, nan)	100 (99.4, 100.0)	(nan, nan)	100 (99.4, 100.0)
	Z-value	(nan, nan)	99.2 (98.1, 99.6)	0 (0.0, 43.4)	100 (99.4, 100.0)
	Local outlier factor	(nan, nan)	96.1 (94.2, 97.3)	0 (0.0, 13.8)	100 (99.3, 100.0)
	Cutoff	(nan, nan)	100 (99.4, 100.0)	(nan, nan)	100 (99.4, 100.0)
	Neural Network				
Diastolic, NIBP	IQR	(nan, nan)	100 (99.4, 100.0)	(nan, nan)	100 (99.4, 100.0)
	Z-value	(nan, nan)	98.9 (97.7, 99.4)	0 (0.0, 35.4)	100 (99.4, 100.0)
	Local outlier factor	(nan, nan)	96.1 (94.2, 97.3)	0 (0.0, 13.8)	100 (99.3, 100.0)
	Cutoff	(nan, nan)	100 (99.4, 100.0)	(nan, nan)	100 (99.4, 100.0)
	Neural Network				
MAP, NIBP	IQR	0 (0.0, 65.8)	100 (99.4, 100.0)	(nan, nan)	99.7 (98.8, 99.9)
	Z-value *	50 (9.5, 90.5)	99 (97.9, 99.5)	14.3 (2.6, 51.3)	99.8 (99.1, 100.0)

	Local outlier factor *	50 (9.5, 90.5)	96.2 (94.4, 97.5)	4.2 (0.7, 20.2)	99.8 (99.0, 100.0)
	Cutoff	0 (0.0, 65.8)	100 (99.4, 100.0)	(nan, nan)	99.7 (98.8, 99.9)
	Neural Network				

Table 4 shows the sensitivity, specificity, positive predictive value, and negative predictive value of all artifact detection algorithms in the ICU. CI95: 95% confidence interval, OR: operating room, PPV: positive predictive value, NPV: negative predictive value, MAP: mean arterial pressure, Systolic: systolic blood pressure, Diastolic: diastolic blood pressure, IBP: invasive blood pressure, NIBP: non-invasive blood pressure, IQR: interquartile range, Neural Network: long-short term memory. Note that too little information was available to train the neural net for non-invasive blood pressure. All cells with a sensitivity above 70% and a specificity above 95% are marked in bold. Asterisks mark methods with the highest sensitivity, specificity was >96% in all marked methods.

Table 5: Values used for Sensitivity Analysis

Threshold	95% CI	Physiologic	Worrying	Urgent
HR [/min]	40-117	50-80	30-140	25-200
Systolic blood pressure[mmHg]	70-170	90-140	60-180	40-240
MAP [mmHg]	45-113	60-80	50-100	35-140
Diastolic blood pressure [mmHg]	28-90	60-90	40-90	20-130
Temp [°C]	34-39.9	36-38	33-41	30-42
SpO2 [%]	95-100	90-100	80-100	50-100
etCO2 [mmHg]	27-47	35-45	30-55	25-80

Table 5 shows the Thresholds used for cut-off method in the sensitivity analysis. Values outside the shown limits were defined as artifacts. HR: Heartrate, Systolic: Systolic blood pressure, MAP: Mean arterial blood pressure, Diastolic: diastolic blood pressure, Temp: Temperature, SpO2: oxygen saturation, etCO2: capnometry, 95% CI: values inside a 95% CI in the dataset, Physiologic, Worrying, Urgent: definitions as in the text.

Figure 1