Lavinia Goldermann*, Simon Fonck, Lena Olivier, Sebastian Fritsch, and André Stollenwerk

The Influence of Human Annotation on CNN Performance for Anomaly Detection in ICU Data

https://doi.org/10.1515/cdbme-2025-0192

Abstract: Deep learning methods are increasingly used in clinical artificial intelligence (AI) research, including for detecting anomalies in intensive care data. However, their evaluation often depends on human annotations, which can vary in quality and consistency. In this study, we analyse the effect of annotation variability on the performance of DeepAnT, an unsupervised convolutional neural network for anomaly detection (AD). Using intensive care time-series data from 38 patients for training and six patients separately annotated for evaluation, we compare F1 scores based on two independent physician annotations. Our results show differences in model performance across different vital parameters, between patients, and especially between annotators evaluating the same data. These findings indicate that human labelling has a measurable impact on the perceived performance of the AD algorithm. Structured labelling protocols may be beneficial for achieving more consistent and reliable evaluations.

Keywords: Artificial Intelligence, Data Quality, Human Annotations, Anomaly Detection

1 Introduction

The use of artificial intelligence (AI) in clinical settings has increased substantially in recent years, particularly in intensive care units (ICU) where continuous monitoring generates large amounts of time-series data. An application of AI in this context is anomaly detection (AD), the automated identification of abnormal data points that may indicate sensor malfunction, data artefacts or clinically relevant events. Deep learning models have shown great potential in detecting such anomalies, especially in univariate or multivariate vital parameter recordings [1]. However, AD models are often based on human-created labels for assessment and, depending on the model, also for training. In the clinical setting, annotations are typically created by physicians can distinguish between e.g. physiological abnormalities and technical artefacts such as sensor dislocation or signal loss. These annotation tasks are not only time-consuming, but also prone to subjective interpretation, differing levels of clinical experience, and the absence of standardised labelling protocols [2]. As a result, both inter-individual inconsistencies (i.e. differences between annotators) and intra-individual inconsistencies (e.g. variations in how a single annotator labels similar data) can affect the quality and consistency of the resulting labels and thus the evaluation of AD models. While previous research has investigated various algorithmic approaches to detecting anomalies in time series data [3], the influence of the quality of human annotations on model performance has not yet been sufficiently investigated. This is particularly problematic when such annotations serve as the ground truth for performance metrics such as precision, recall or F1 score.

In this study, we aim to assess the impact of annotations variability on the performance of a deep learning-based AD model, DeepAnT [4], applied to real patient data from intensive care units. We investigate how the model's F1 scores differ when evaluated on annotations from two independent physicians, and how these deviations are reflected in different patients and vital parameters.

2 Related Work

For a better understanding of how our study contributes to this area, we want to review the existing work related to AD in ICU data, as well as research on the impact and variability of human annotations in clinical AI. AD is a widely researched topic and is applied in many domains. Chandola et al. describe various applications in which AD is used [5]. These include intrusion detection to monitor network traffic and im-

***Corresponding author: Lavinia Goldermann,** Embedded Software (Informatik 11), RWTH Aachen University, Templergraben 55, D-52064 Aachen, Germany, e-mail: goldermann@embedded.rwth-aachen.de
Simon Fonck, André Stollenwerk, Embedded Software (Informatik 11), RWTH Aachen University, Aachen, Germany
Lena Olivier, Neonatology Section of the Department of Paediatric and Adolescent Medicine, RWTH Aachen University Hospital, Aachen, Germany
Sebastian Fritsch, Jülich Supercomputing Centre (JSC), Forschungscentrum Jülich, 52428 Jülich, Germany
Simon Fonck, Sebastian Fritsch, André Stollenwerk, Center for Advanced Simulation and Analytics (CASA), Forschungszentrum Jülich, 52428 Jülich, Germany

prove cyber security, fault detection in safety-critical systems, and credit card fraud detection. In addition, numerous summaries of state-of-the-art AD algorithms for univariate and multivariate time series data are presented by other authors. Specifically, Zamanzadeh et al. provide an extensive review of AD models, focusing on architectures, methodological approaches, and the availability of implementations [3]. Munir et al. compared 13 AD methods [1] using two benchmark data sets: Yahoo Webscope [6] and NAB [7], which contain time series from areas such as transport and cloud computing. Among the evaluated methods, the deep learning models FuseAD [8] and DeepAnT [4] performed the best. Although the use of AD has been growing in various fields, its application in the medical field, especially in ICU datasets, remains limited. One of the few examples is presented by Salem et al., who proposed a kernel density-based distance measure to detect point anomalies in multivariate medical time series without requiring prior knowledge of future data points [9]. Their method demonstrated high performance when evaluated on the MIMIC-III dataset [10], achieving a sensitivity of 100% and specificity of 94.5%. However, their approach does not rely on deep learning techniques, nor does it incorporate physician-annotated ground truth. Based on the promising results of Munir et al. and the current lack of deep learning-based AD studies on ICU datasets, we chose to apply DeepAnT to this domain. To our knowledge, this represents one of the first attempts to analyse ICU time series data using DeepAnT, with a focus on how annotation quality affects model performance. Annotation quality is especially important for clinical AI, as shown by Sylolypavan et al. [2]. In their study, 11 ICU physicians annotated clinical cases individually, and separate machine learning models were trained on each annotation set. The results showed significant variation in model outputs that was solely due to labelling inconsistencies. This work illustrated the challenges of assuming that a single set of annotations represents a definitive truth and called for more structured and transparent annotation processes.

Given the findings by Sylolypavan et al., we aim to explore whether similar effects can be observed when using deep learning-based AD on ICU time series data. Specifically, we investigate how differing annotations by two physicians influence the evaluation of DeepAnT, and to what extent these inconsistencies affect the interpretability and trustworthiness of model outputs.

3 Methods

The data used in this study were provided by the RWTH Aachen University Hospital (UKA). It consists of ICU records collected between 2007 and 2019. For each patient, several vital parameters were continuously monitored during their stay in the ICU. From the available parameters, we selected eight vital parameters (as listed in Table 1) for further analysis. These parameters were chosen because of their high data density, making them particularly suitable for data-driven AD approaches. To identify anomalies in the time-series data, we used DeepAnT [4], designed for AD in both univariate and multivariate time series data. The model predicts future values based on past input windows. One of DeepAnT's key strengths is its robustness: it is able to learn effectively even when the training data contains anomalies, as these do not significantly affect the model's learning process. This is particularly useful considering that many AD approaches rely on learning a representation of the normal state of the data in order to identify deviations. Furthermore, it can produce reliable results even when trained on relatively small datasets [4].

In our study, DeepAnT was trained univariately with data from 38 patients. This means that a separate model was trained for each of the eight selected vital signs. The total number of training data points for each model is listed in Table 1. For the validation phase, we used data from an additional six patients. Both datasets were independently annotated by two physicians (physician A and physician B). The annotations focused on identifying non-physiological anomalies such as sensor failures, dislocated sensors and other artefacts that could affect the quality of the data. We calculated the F1 score to evaluate the correspondence between the anomalies detected by DeepAnT and the annotations created by the physician. This metric provides a balanced measure of precision and recall and allows a comparison between the model result and the 'ground truth' created by humans.

Tab. 1: Overview of vital parameters, their acronyms and total number of training data points from 38 patients for the selected CNN.

Vital Parameter	Acronym	Number of training data points
Central Venous Pressure	CVP	117,014
Heartrate	HR	2,369,220
Respiratory Rate	RR	1,816,632
Oxygen Saturation	SaO2	2,052,516
Temperature	TempC	217,564
Diastolic Pressure	pDi	348,813
Mean Arterial Pressure	pMe	350,517
Systolic Pressure	pSy	348,807

4 Results

The evaluation of DeepAnT's performance was conducted on data from six ICU patients, using a physician's annotations as

a reference. Table 2 presents the F1 scores achieved by DeepAnT across all eight selected vital parameters for each patient. For this evaluation, annotations from *physician A* were used as the basis for comparison. The table shows a wide range of F1 scores, with notable differences not only between vital parameters, but also between patients. In cases where no F1 score is reported, the parameter was not recorded for that patient.

A closer look at the parameter-wise performance shows that *TempC* consistently achieved high F1 scores, ranging from 60.7 to 100% across all six patients. This suggests that DeepAnT is particularly effective at detecting anomalies in temperature data. In contrast, other parameters showed more variable results.

Patient-wise variation in model performance was also observed. For example, Patient 2 consistently achieved F1 scores above 72.3%, with *SaO2* reaching the maximum score of 100%. This contrasts with other patients where scores were significantly lower in some cases and more heterogeneous across the different parameters.

To further investigate the impact of human annotation on model performance, we repeated the analysis for selected patients using the annotations provided by *physician B*. These results are summarised in Table 3. In addition to the F1 scores, the table also includes the number of anomalies annotated by each physician for every vital parameter. Here, we observed significant differences in the F1 values for the same data when we compared the model results with the two independent annotations. These variations can be partly explained by large differences in the number of labelled anomalies. For instance, in Table 3 (a), physician A annotated 29 anomalies in the parameter *pSy*, while physician B annotated 68. In contrast, for patient 2 in the parameter *SaO2*, both physicians identified the same single anomaly, resulting in perfect agreement and a maximal F1 score.

These findings show the importance of annotation consistency and the influence that subjective labelling can have on the evaluation of AD algorithms.

Tab. 2: F1 Scores of DeepAnT for each patient and vital parameter (evaluated with Physician A's annotations).

	Pat. 1	Pat. 2	Pat. 3	Pat. 4	Pat. 5	Pat. 6
CVP	–	–	**0.944**	0.250	–	**0.750**
HR	0.244	**0.800**	0.500	0.385	0.440	0.326
RR	0.333	**0.723**	0.540	0.089	0.540	0.491
SaO2	0.279	**1.000**	0.407	**0.800**	0.407	0.475
TempC	**0.607**	**0.920**	**0.738**	**1.000**	**0.611**	**0.783**
pDi	0.520	**0.889**	**0.720**	0.508	**0.947**	**0.628**
pMe	0.472	**0.857**	**0.824**	0.552	**0.762**	0.434
pSy	0.419	**0.750**	0.533	0.359	**0.543**	0.387

Tab. 3: Comparison of F1 scores of and the number of annotated anomalies between physician A and B for patient 1 (a) and 2 (b).

(a) Patient 1

	HR	RR	SaO2	TempC	pDi	pMe	pSy
A	**0.244**	**0.333**	**0.279**	0.607	0.520	0.472	**0.419**
	35	21	30	61	32	43	29
B	**0.417**	**0.167**	**0.480**	0.656	0.500	0.488	**0.275**
	22	15	12	51	25	28	68

(b) Patient 2

	HR	RR	SaO2	TempC	pDi	pMe	pSy
A	**0.800**	0.723	1.000	0.920	0.889	**0.857**	**0.750**
	12	39	1	147	14	33	15
B	**0.570**	0.600	1.000	0.916	0.636	**0.471**	**0.571**
	2	6	1	128	14	9	6

5 Discussion

The results presented in this study show that the evaluation of the AD performance of DeepAnT is highly dependent on the quality and consistency of the underlying human annotations. For the six ICU patients analysed, we observed significant variations in F1 values depending on the vital parameters and the specific patient data set. For example, parameters such as *TempC* consistently achieved high F1 scores, while others showed much greater variability. Similarly, some patient data sets consistently performed better, suggesting differences in data quality or clarity of annotations.

In addition, comparing results based on two independent physician annotations for the same patient showed considerable differences. In some cases, the F1 scores differed significantly between the two sets of annotations, despite the underlying data remaining unchanged. Similar to the study by Sylolypavan et al. [2], we were able to demonstrate the influence that subjective human annotations can have on model performance and point to the need for structured or standardised annotation protocols. Without such consistency, comparisons between models or general conclusions about performance become questionable, especially if the evaluation is based on potentially erroneous or inconsistent labels.

These results highlight a problem in clinical AI research: while attention is paid to algorithmic improvements, the foundations of the evaluation, namely the annotation process, are often insufficiently investigated. For AD models to be meaningfully validated, the annotation process must be approached with the same care as the model development itself. This includes establishing an initial consensus among annotators regarding annotation criteria and thresholds for anomalies. Ide-

ally, the annotation process should begin with a learning phase, in which annotators align their understanding and interpretations before independently labelling datasets. If adjustments or refinements to the annotation strategy are required over time, these should be made through clearly documented and traceable iterations to avoid subsequent bias.

5.1 Limitations

One of the main limitations of this study is the small sample size of the datasets that were tested. Due to the limited availability of annotated ICU data, only six patient data sets were available for evaluation, which limits the generalisability of the results. Furthermore, only two annotators were involved in the labelling process. To better assess the agreement between annotators and to reduce individual bias, at least three annotators would be preferable, enabling pairwise comparisons such as B and C against A. Ideally, even more annotators should be involved to gain a better understanding of the variability of the annotation.

In addition, our analysis focused exclusively on the DeepAnT architecture. While this model proved to be viable in the context of univariate AD, it remains unclear whether the observed annotation sensitivity can be transferred to other architectures or multivariate settings. Future work should therefore include a wider range of models and more extensive data sets.

Another limitation is the unavailability of the intensive care data used in this study for external validation. Although we intend to make the annotations available for public datasets, this has not yet been realised.

6 Conclusions

In this study, we investigated the influence of human annotation on the performance evaluation of a deep learning-based AD algorithm on ICU time series data. By comparing the F1 scores of DeepAnT across different patients, vital parameters and independent physician annotations, we demonstrated that the quality and consistency of the labels have a significant impact on the resulting performance metrics.

Our results highlight an underestimated aspect of clinical research. The evaluation of model performance is only as reliable as the annotations on which it is based. Differences between annotators can lead to significant variations in perceived model quality, directly affecting model benchmarking, validation, and potential clinical use. To improve the reliability of such assessments, structured and standardised annotation protocols could be created. The involvement of a larger number of annotators and the implementation of consensus-building strategies can help to reduce individual bias. Furthermore, AI-supported labelling approaches such as active learning or semi-supervised methods could serve as powerful tools to support physicians and improve labelling consistency while reducing the manual workload.

Future work should aim to increase both the number of annotated data sets and the number of AD models examined. Furthermore, making annotated data sets publicly available could increase reproducibility and enable the development of a common annotation structure.

Author Statement

Research funding: The author state no funding involved. Conflict of interest: Authors state no conflict of interest. Informed consent: Informed consent has been obtained from all individuals included in this study. Ethical approval: not applicable, as secondary research data was used.

References

[1] Munir M., Chattha M. A., Dengel A., and Ahmed S. A comparative analysis of traditional and deep learning-based anomaly detection methods for streaming data. In 2019 18th IEEE international conference on machine learning and applications (ICMLA), pages 561–566. IEEE, 2019.

[2] Sylolypavan A., Sleeman D., Wu H., Sim M. The impact of inconsistent human annotations on AI driven clinical decision making. NPJ Digit Med. 2023

[3] Darban Z. Z., Webb G. I., Pan S., Aggarwal C., and Salehi M. Deep learning for time series anomaly detection: A survey. ACM Computing Surveys, 57(1), 1-42, 2024.

[4] Munir M., Siddiqui S. A., Dengel A., and Ahmed S. Deepant: A deep learning approach for unsupervised anomaly detection in time series. IEEE Access, 7:1991–2005, 2019.

[5] Chandola V., Banerjee A., and Kumar V. Anomaly detection: A survey. ACM Comput. Surv., 41, 07 2009.

[6] Laptev N., Amizadeh S., and Flint I. Generic and scalable framework for automated time-series anomaly detection. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 1939–1947. ACM, 2015.

[7] Lavin A. and Ahmad S. Evaluating real-time anomaly detection algorithms–the numenta anomaly benchmark. In 2015 IEEE 14th international conference on machine learning and applications (ICMLA), pages 38–44. IEEE, 2015.

[8] Munir M., Siddiqui S. A., Chattha M. A., Dengel A., and Ahmed S. Fusead: unsupervised anomaly detection in streaming sensors data by fusing statistical and deep learning models. Sensors, 19(11):2451, 2019.

[9] Salem O., Liu Y., and Mehaoua A. Anomaly detection in medical wireless sensor networks. Journal of Computing Science and Engineering, 7(4):272–284, 2013.

[10] Johnson A. E. W., Pollard T. J., Shen L., Lehman L. wei H., Feng M., Ghassemi M., Moody B., Szolovits P., Celi L. A., and Mark R. G. MIMIC-III, a freely accessible critical care database. Scientific Data, 3(1), May 2016.