Towards Design Principles for a Real-Time Anomaly Detection Algorithm Benchmark Suited to Industrie 4.0 Streaming Data

Philip Stahmann
University of Osnabrueck
philip.stahmann@uni-osnabrueck.de

Prof. Dr.-Ing. Bodo Rieger
University of Osnabrueck
bodo.rieger@uni-osnabrueck.de

Abstract

The vision of Industrie 4.0 includes the automated reduction of anomalies in flexibly combined production machine groups up to a zero-failure ideal. Algorithmic real-time detection of production anomalies may build the basis for machine self-diagnosis and self-repair during production. Several real-time anomaly detection algorithms appeared in recent years. However, different algorithms applied to the same data may result in contradictory detections. Thus, real-time anomaly detection in Industrie 4.0 machine groups may require a benchmark ranking for algorithms to increase detection results' reliability. This paper makes a qualitative research contribution based on ten expert interviews to find design principles for such a benchmark ranking. The experts were interviewed on three categories, namely timeliness, thresholds and qualitative classification. The study's results can be used as groundwork for a prototypical implementation of a benchmark.

1. Introduction

In 2011, the vision of Industrie 4.0 was initiated in Germany to characterize the production paradigm of the future [11]. An essential target of Industrie 4.0 is the autonomous control of production machines [17]. To this end, machines shall be equipped with a range of so-called self-X competencies [5]. Among these, self-diagnosis and self-repair include the autonomous detection and elimination of anomalies during production processes. Since about 2013, the number of publications on automated real-time anomaly detection has been increasing continuously [18]. The year 2021 ushers in the second decade of the Industrie 4.0 vision and self-diagnosis and self-repair competencies continue to form essential components [2].

At the same time, the ongoing digitalization of production environments in the context of Industrie 4.0 supposedly increases production flexibility [11]. Using technological capabilities to combine single production machines to temporary groups performing tasks jointly shall enable meeting individual customer needs and lead to competitive advantages. However, anomalous behavior in one machine may propagate to connected group members [6]. Finding anomalies' roots is more complex in flexible configurations consisting of multiple machines [19]. Moreover, supposedly inconsequential anomalies that might remain unnoticed by experts may significantly affect other parts of the production ensemble. This rising complexity may require automated, algorithmic anomaly detection in production even more [10].

To monitor production with regards to anomalies, companies increasingly deploy sensor technology [8]. Sensors enable real-time views on production by constantly emitting data streams [9]. There are several domain independent openly accessible algorithms that aim to detect anomalies in such streams [16]. Usually, several algorithms are deployed in parallel to determine anomalies as there is no single algorithm that fits all scenarios best [18]. To the best of our knowledge no Industrie 4.0 specific factors have been defined to support the selection of fitting real-time anomaly detection algorithms. The deployment of different algorithms to the same data to determine production anomalies may lead to contradictory results. These contradictions put decision makers in a dilemma. On the one hand, they may decide that an anomaly exists when there is an indication by one of multiple algorithms and invest work in attempting to correct the anomaly, at the risk of reacting to false alarms. On the other hand, decision makers may decide to fix anomalies only on the indication of multiple algorithms, which can lead to smaller anomalies being overlooked and propagated in the machine groups. To preventively avoid such dilemmas, it is necessary to understand which factors may support algorithm benchmarking, i.e. evaluation and prioritization, and how to implement them. On this basis, we formulate the following research question: *Which Industrie 4.0 specific design principles can be defined to benchmark real-time anomaly detection algorithms?* To find design principles, we conducted qualitative interviews with industry experts to get an understanding on what matters to them regarding

URI: https://hdl.handle.net/10125/80106
978-0-9981331-5-7

anomaly detection evaluation. We restricted the scope of investigation to the three categories timeliness, threshold setting and qualitative classification as these are discussed in related studies [7, 13, 18].

The paper proceeds as follows. Section two outlines evaluation categories for anomaly detection algorithms in Industrie 4.0. Section three details the qualitative research approach taken to find answers to the research question. Sections four and five aggregate the research results and provide a discussion and conclusion.

2. Evaluation categories for real-time anomaly detection in Industrie 4.0

Several different open source real-time anomaly detection algorithms have been developed in recent years [16]. In general, all of them underlie the assumption that the vast majority of generated data adheres to expectation, whereas anomalies are exceptional [4]. Yet, the application of different algorithms leads to different results. Figure 1 exemplarily shows the results of five state-of-the-art algorithms from [16] applied to a real data set of machine temperature measurements over a span of four days. The differences in detection results make it difficult to interpret where anomalies actually occur. For this reason, [13] have developed a benchmark to prioritize real-time anomaly detection algorithms. However, in the benchmark, algorithms are evaluated without further context. In order to evaluate algorithms specifically for the Industrie 4.0 context, design principles for a specific benchmark are required. Context-specificity results from mechanisms used to analyze and evaluate production data regarding adherence to expectation. In related studies, three categories for design principles are identified regarding real-time anomaly detection evaluation, namely timeliness, threshold setting and qualitative anomaly assessment as presented in the following [7, 13, 18].

The first category refers to the time between anomaly occurrence, detection and notification. According to [13], timeliness is of major importance as optimal algorithms are supposed to detect anomalies as early as possible, so that countermeasures against anomalies can be initiated as soon as possible. The structured literature review presented in [18] identifies requirements for real-time anomaly detection in Industrie 4.0. The majority of analytical requirements such as fast data preparation emphasizes the importance of timeliness. The possibility of analyzing various influences in production in real-time is given by the progress in sensor technology [17]. Smart sensors are able to measure and communicate various signals for analysis in real time. Furthermore, there are different types of notifications, such as audible alarms or visual pop-up messages. [1] implement a middleware in which notifications should reach the correct addressee as quickly as possible. For this purpose, they minimize the time between anomaly detection and notification.

Threshold setting is an intensively discussed topic in practical applications for anomaly detection in Industrie 4.0 [12, 18]. Thresholds are upper and lower bounds normal values shall not trespass [3]. Each data occurrence outside these thresholds is declared as anomalous. Threshold values may change during analysis as they may have to adapt to production conditions [13, 16]. For example, the maximum expected temperature of an engine shortly after start may be initially lower than during full operation. A key argument for considering threshold adherence in anomaly detection is that it can be easily controlled at low cost. In addition, appropriate measures can be prepared for cases of threshold violations. [12].

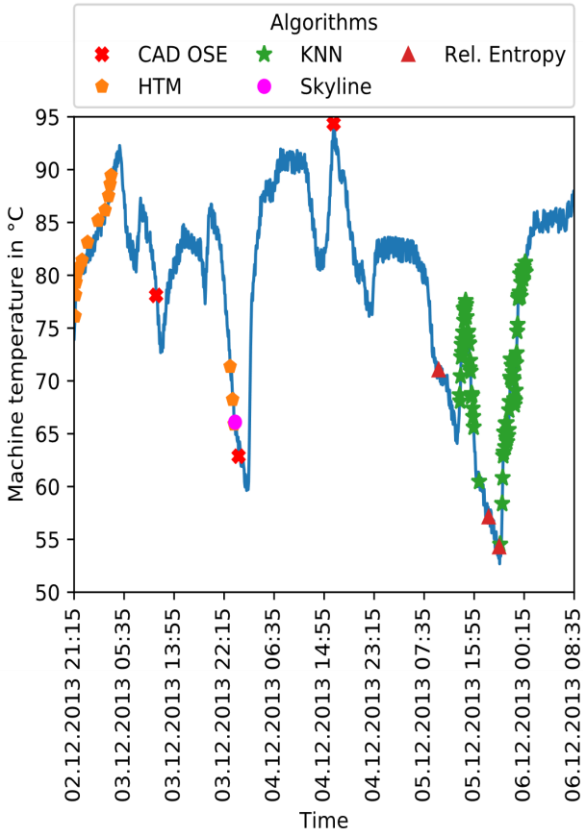

Figure 1. Different anomaly detection results per algorithm applied to an extract from a real-world machine temperature data set

The real-time anomaly detection algorithm benchmark in [13] considers whether an anomaly is detected by an algorithm or not. This means that the benchmark restricts to the detection of true and false positive and negative anomaly detections. A qualitative classification, e.g. whether certain anomalies should be

detected with higher priority than others is not included. However, research shows the need for qualitative anomaly assessment [7, 14]. [7] develop requirements for data sets that can be used for benchmarking anomaly detection algorithms. According to the authors, each instance in data sets should be assigned to meaningful, qualitative categories. Examples of such categories are detection difficulty or anomaly impact. In this regard, anomaly detection algorithms may be benchmarked according to their capability to detect anomalies belonging to the most difficult or impactful class.

3. Methodology

Ten qualitative, semi-structured interviews with industry experts have been conducted to get impressions of how real-time anomaly detection in terms of timeliness, threshold setting and qualitative anomaly assessment is evaluated in practice. The main goal is to formulate design principles for a real-time anomaly detection algorithm benchmark based on these impressions. The interviews were conducted individually in a period of three months and correspond to the guidelines proposed by [4].

They took about 30 minutes each. The language of choice was German as all participants were native speakers. The results were translated to English after all interviews. At the beginning of each interview, the interviewees were first asked about the size of the company they work for and their role in the company, as well as the length of their practical experience in real-time anomaly detection in Industrie 4.0 production environments. Table 1 shows information on the interviewees such as their own role description.

Machine operators are directly involved in the detection of anomalies through their daily involvement in production processes. They do not calibrate or parametrize anomaly detection mechanisms themselves. However, they read out and interpret anomaly detection results during production to start appropriate countermeasures. Therefore, they contribute experience in detecting and handling anomalies to our study. Manufacturing engineers as well as project engineers consider anomaly detection mechanisms in production planning. To this end, they need to consider potential anomalies' effects on production. Information on potential anomalies may result from prior experience. Table 2 provides details on the interviewees' companies. All interviewees are German, yet, we emphasize that nine of ten interviewees work in internationally operating companies, so that results are not necessarily restricted to German manufacturing contexts. All interviewees have at least two years of practical experience with real-time anomaly detection in production environments in the indicated fields of occupation. Both industrial machine manufacturing companies focus on customer individual production, the others produce for the general market. After gathering the information presented in tables 1 and 2, the interviewees were asked on the three categories, where the purpose of each category was briefly explained before the questions.

Table 1: Information on interviewees

Occupation	Role description	No.
Machine operator	Responsible for operational setup and maintenance of production machine groups as directed by machine manufacturers and manufacturing engineers.	5
Manu-facturing engineer	Responsible for medium to long-term planning and maintenance of production machine groups.	3
Project engineer	Scheduling and use of machine groups in different, partially dependent production programs.	2

Table 2: Information on companies

Industry	Operated country	No.
Machine spare parts manufacturing	Worldwide (1), Germany (2)	3
Vehicle manufacturing	Worldwide	4
Industrial machine manufacturing	Germany, Netherlands	2
Industrial production automatization	Germany	1

Before starting with the first category, the interviewees were advised that the questions are only guidelines and that a free, open dialogue on the topic is desired [4]. Table 3 shows the questions from the semi-structured questionnaire. All questions are generic to enhance free, unbiased answers.

The first question regarding timeliness is supposed to deliver answers on time robustness of anomaly detection and notification. The hint was given that the question implied that anomalies may be detected and reported late or early after their occurrence. Answers to this category took about eight minutes on average, all interviewees responded to the question.

Regarding threshold setting as second category two questions were asked. The first one is an open question on the consideration of thresholds in real-time anomaly detection. The question was explicitly restricted to thresholds used for distinction of normal and anomalous production values to prevent confusion with other production-related thresholds, such as minimum number of employees for production supervision. All interviewees responded to that question. Secondly, a narrower question on the measure for threshold value definition was asked. Due to the lack of information, two interviewees could not provide an answer on the second question. Answering the questions from the second category took about eleven minutes.

Table 3: Semi-structured questionnaire

	Timeliness
Q1:	In how far should real-time anomaly detection evaluation consider the timely detection and notification of potential anomalies in a data sequence?
Hint:	"Timely detection and notification" may mean both early or late detection and notification after anomalous occurrences.
	Threshold setting
Q2:	In how far should real-time anomaly detection evaluation consider thresholds?
Hint:	The question is limited to thresholds that distinguish expectable production values from anomalies.
Q3:	On what basis do you define thresholds for real-time anomaly detection?
	Qualitative classification
Q4:	In how far should real-time anomaly detection evaluation differentiate anomalies qualitatively?
Hint:	Qualitative anomaly differentiation refers to potential differences regarding kinds of anomalies.

In the last category, the interviewees were asked an open question on considerations of qualitative differences among anomalies in their companies. As most respondents had difficulty understanding the question, they were given the hint that qualitative differences may refer to different kinds of anomalies. After that, all interviewees answered the question in about nine minutes on average.

After all interviews, the answers were analyzed and reduced following the steps in [15]. In line with the "paraphrasing" step, statements were classified according to the question categories. This was necessary because some interviewees returned to previous questions in the course of the interview. Furthermore, the answers were examined for digressions, which were subsequently removed. According to the "generalization" step, the essential content of the statements was concisely formulated in short phrases. Finally, two reductions were made. The reductions included a summary and an aggregation of the short phrases and the subsequent induction of design principles [15].

4. Results

Table 4 shows the results of the qualitative interviews. The second column contains a list of phrases that were mentioned by more than five interviewees. Column three contains the design principles as generic statements resulting after two reductions following [15]. Column four shows the number of interviewees who made statements during the interview that can be subsumed under the design principles. With regard to timeliness, most interviewees stated that they are aware that the time lag between anomaly occurrence, detection and notification is documented. The time margin when an anomaly is reported is adjustable depending on the machine group and production project. The advantage of reporting as early as possible after occurrence is that an immediate reaction is possible. However, most interviewees state that an early report of the anomaly also means that no further, longer analysis takes place and that the consequences of the potential anomaly are not considered. I.e., the possibility that a potential anomaly is inconsequential or of low relevance and thus negligible is discarded when an alarm for an anomalous occurrence is raised as each alarm is taken seriously. As a result, early reports may often lead to negligible or even false alarms. Consequences, such as shutting down production to accurately identify and eliminate the anomaly, may thus be erroneously initiated. Such unnecessary actions have financial consequences, because they delay production. For this reason, eight of the interviewees emphasize the need for time robustness. Specifically, this means that further automated analysis as well as monitoring of consequences should be initiated after the occurrence of potential anomalies, but not directly a notification that an anomaly exists.

All interviewees found that thresholds are the most important indicators of whether a value measured in production is normal or anomalous. In fact, seven of them took longer than a minute to come up with methods that are alternative to thresholds and support delineating anomalies. An example of an alternative method is the visual inspection of pressed metal sheets for cracks.

Table 4: Results from qualitative interviews

Question	Phrases mentioned by more than five interviewees	Design principles after two reductions (following [15])	No.
	Timeliness		
Q1	Time of occurrence and detection is always measured and documented; Time lag between occurrence and notification is measured and documented	**DP1:** Real-time anomaly detection evaluation needs to consider whether anomalies were detected and notified early or late after occurrence.	7
	Immediate detection and notification are advantageous for countermeasures; Delayed notification is sufficient most of the time; Delayed notification is more robust against false alarms; Delayed notification enables longer analysis; Robustness required; False alarms hinder process significantly	**DP2:** Real-time anomaly detection evaluation needs to consider that raising early anomaly alarms might increase false alarm rates. Therefore, real-time anomaly detection evaluation needs to consider a certain robustness.	8
	Threshold setting		
Q2	Differentiation happens almost always on the basis of thresholds; Very few alternatives to thresholds are used	**DP3:** Real-time anomaly detection evaluation needs to consider thresholds, as these are the most frequent mechanism for anomaly detection in production.	10
Q3	Not documented experts' experience is used; Expert consultation is necessary; Ad-hoc intuition is used; All production processes are simulated; Manufacturer provides threshold values; Strong dependence on material composition; Material-specific information is used	**DP4:** Real-time anomaly detection evaluation needs to consider personal experience from production step specific experts as well as simulation and statistics on materials to set fixed and dynamic thresholds.	7
	Qualitative classification		
Q4	Qualitative anomaly classification builds on impact calculation;	**DP5:** Real-time anomaly detection evaluation needs to consider systematic classification of anomalies according to their impact.	7
	Impact calculated based on time of machine standstill; Impact calculated based on number of influenced production machines; Impact calculated based on work effort of countermeasures; Impact calculated based on financial effort of countermeasures;	**DP6:** Real-time anomaly detection evaluation needs to consider time, cost and intensity requirements of countermeasures to determine anomalies' impact and thus classification.	8

Two different types of thresholds were described, fixed thresholds apply unchanged throughout the production period. An example is the target weight of produced workpieces. Other threshold values are dynamic over time, because they depend on changing production variables. For example, the temperature of a press machine for metal sheets changes with the number of pressed sheets per minute due to friction. If the temperature of the sheets is now measured, threshold values that delimit whether a temperature value is too high or too low must be adjusted accordingly.

The interviewees mentioned three different sources for the determination of threshold values. The personal experience of machine operators and especially machine engineers was mentioned by six interviewees. Personal experience is not documented. Corrective actions can also be based on situational intuition. Furthermore, the simulation of production processes helps to determine threshold values before the start of production. Thirdly,

statistics on the material of the machine and the workpiece support the determination of threshold values. These are either taken from empirical sources or are already given by the manufacturer of the machines.

Seven interviewees were aware of the qualitative classification of anomalies after the hint from table 3 was given. Three mentioned a classification they were aware of into negligible, relevant, strong, and critical anomalies. The classification depends on the difference in the impact of an anomaly. Examples of anomaly impact include the number of production machines affected or machine downtime required to eliminate an anomaly.

Furthermore, time and effort required for countermeasures influence the classification. The interviewees mentioned financial impact as most important for classifying anomalies. This includes not only the wear and tear on parts due to anomalies, but also the cost of countermeasures to correct the anomalies. The more expensive the impact of an anomaly, the more important it is to detect or prevent it.

5. Discussion and conclusion

Machine self-diagnosis and self-repair competencies are important building blocks for fulfilling the Industrie 4.0 vision. Various algorithms can support these competencies through real-time anomaly detection. However, Industrie 4.0 decision makers face the difficulty of identifying the most suitable algorithm. In this study, qualitative interviews were conducted with ten industry experts to inductively derive design principles for a benchmark that supports evaluation and prioritization of real-time anomaly detection algorithms. Building on previous research, the interviews were divided into the categories timeliness, threshold setting, and qualitative classification. This study contributes with practical in-depth knowledge extending the state-of-the-art in these three categories. As shown in table 1, the interviewees contribute from different perspectives due to the different roles they have in production companies. Derived from the repetition of certain phrases in different interviews, six design principles could be formulated. Table 4 shows the results of the survey.

Regarding the results for the category timeliness there is a clear indication that the time between the occurrence, detection and reporting of an anomaly is measured and documented. However, there is no precise specification of what is early or late. It is therefore hardly possible to draw a clear conclusion as to when an anomaly is detected or reported too early or too late. The interviewees confirm the relevance of threshold setting for anomaly detection. Simulation, statistics on production material composition, such as maximum allowed heat for certain materials, and personal experience result as useful sources for threshold value determination. The reliance on personal experience makes the exact procedure difficult to define scientifically, since personal experience and intuition depend on the decision maker and the situation. The results regarding qualitative classification essentially contribute with the knowledge that the impact of anomalies is relevant for their classification. However, concrete classes could not be found with majority among the experts. Additionally, seven experts emphasized the time and labor intensity of labeling each anomaly in terms of its qualitative classification to obtain a data set that shows which algorithm performs best in this category.

Despite adherence to methodological guidelines from [4] and [15] our study is not free from limitations. The interview's questions are not open-ended, so that responses are limited to the three categories previously identified from literature. Also, it can be criticized that the formulation of design principles is subjective. In order to reduce subjectivity, phrases mentioned by more than five interviewees were included in table 4. Moreover, the results are difficult to generalize because only ten experts from industry were interviewed. However, the identification of experts is complicated by the fact that the topic requires very specific practical experience.

The results contribute an indication of what to consider when evaluating anomaly detection algorithms. They can be used as groundwork for future studies. These could e.g. extend the identified design principles on the basis of open-ended questions. Another possible future study may be the prototypical implementation of a benchmark that prioritizes and evaluates algorithms for real-time anomaly detection in the context of Industrie 4.0.

Complementary to the results, six interviewees stressed that not all anomalous occurrences are real-time detectable in their companies. Regular maintenance intervals serve the need to analyze past anomalies that were not detected upon occurrence. In this context, they mentioned that they perceive the focus on real-time as very specific. However, all interviewees consider the evaluation of real-time anomaly detection algorithms relevant to leverage potentials in production monitoring and control.

6. References

[1] Ali, M. I., Patel, P., and Breslin, J. G. 2019 - 2019. Middleware for Real-Time Event Detection andPredictive Analytics in Smart Manufacturing. In *2019 15th International Conference on Distributed Computing in Sensor Systems (DCOSS)*. IEEE, 370–376. DOI=10.1109/DCOSS.2019.00079.

[2] Balzereit, K. and Niggemann, O. 2020 - 2020. Automated Reconfiguration of Cyber-Physical Production Systems using Satisfiability Modulo Theories. In *2020 IEEE Conference on Industrial Cyberphysical Systems (ICPS)*. IEEE, 461–468. DOI=10.1109/ICPS48405.2020.9274707.

[3] Basu, S. and Meckesheimer, M. 2007. Automatic outlier detection for time series: an application to sensor data. *Knowl Inf Syst* 11, 2, 137–154.

[4] Bell, E., Bryman, A., and Harley, B. 2019. *Business research methods*. Oxford University Press, Oxford.

[5] Cohen, Y. and Singer, G. 2021. A smart process controller framework for Industry 4.0 settings. *J Intell Manuf* 20, 1, 109.

[6] Dafflon, B., Moalla, N., and Ouzrout, Y. 2021. The challenges, approaches, and used techniques of CPS for manufacturing in Industry 4.0: a literature review. *Int J Adv Manuf Technol* 113, 7-8, 2395–2412.

[7] Emmott, A. F., Das, S., Dietterich, T., Fern, A., and Wong, W.-K. 2013. Systematic construction of anomaly detection benchmarks from real data. In *Proceedings of the ACM SIGKDD Workshop on Outlier Detection and Description - ODD '13*. ACM Press, New York, New York, USA, 16–21. DOI=10.1145/2500853.2500858.

[8] Han, S., Gong, T., Nixon, M., Rotvold, E., Lam, K.-Y., and Ramamritham, K. 2018. *RT-DAP: A real-time data analytics platform for large-scale industrial process monitoring and control*.

[9] Henning, S. and Hasselbring, W. 2020. Scalable and Reliable Multi-dimensional Sensor Data Aggregation in Data Streaming Architectures. *Data-Enabled Discov. Appl.* 4, 1, 15:2.

[10] Hsieh, R.-J., Chou, J., and Ho, C.-H. 2019. Unsupervised Online Anomaly Detection on Multivariate Sensing Time Series Data for Smart Manufacturing, 90–97.

[11] Kagermann, H., Wahlster, W., and Helbig, J. 2013. *Umsetzungsempfehlungen für das Zukunftsprojekt Industrie 4.0*. *Abschlussbericht des Arbeitskreises Industrie 4.0*.

[12] Kiangala, K. S. and Wang, Z. 2018. Initiating predictive maintenance for a conveyor motor in a bottling plant using industry 4.0 concepts. *Int J Adv Manuf Technol* 97, 9-12, 3251–3271.

[13] Lavin, A. and Ahmad, S. 2015. Evaluating Real-Time Anomaly Detection Algorithms -- The Numenta Anomaly Benchmark, 38–44.

[14] Maxion, R. A. and Tan, K. 2000. Benchmarking anomaly-based detection systems. In *Proceeding International Conference on Dependable Systems and Networks. DSN 2000*. IEEE Comput. Soc, 623–630. DOI=10.1109/ICDSN.2000.857599.

[15] Mayring, P. 2014. *Qualitative content analysis: theoretical foundation, basic procedures and software solution*. SSOAR, Klagenfurt.

[16] Numenta Anomaly Benchmark. 2020. *Numenta Anomaly Benchmark*. https://github.com/numenta/NAB.

[17] Schütze, A., Helwig, N., and Schneider, T. 2018. Sensors 4.0 – smart sensors and measurement technology enable Industry 4.0. *J. Sens. Sens. Syst.* 7, 1, 359–371.

[18] Stahmann, P. and Rieger, B. 2021. Requirements Identification for Real-Time Anomaly Detection in Industrie 4.0 Machine Groups: A Structured Literature Review. *Proceedings of the 54th Hawaii International Conference on System Sciences*.

[19] Wang, J., Liu, C., Zhu, M., Guo, P., and Hu, Y. 2018. Sensor Data Based System-Level Anomaly Prediction for Smart Manufacturing, 158–165.